Empirical essays on policy interventions in the digital economy

Raphaela Andres

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2024IPPAT016

Thèse de doctorat

Empirical Essays on Policy Interventions in the Digital Economy

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat : Sciences Économiques

Thèse présentée et soutenue à Palaiseau, le 27 mai 2024, par

RAPHAELA ANDRES

Composition du Jury :

Marc Bourreau Professeur d'Économie, Télécom Paris, Institut Polytechnique de Paris	Président/Examinateur
Grazia Cecere Professeure d'Économie, IMT Business School	Rapporteuse
Fabrice Le Guel Professeur d'Économie, Université Paris-Saclay	Rapporteur
Irene Bertschek Professeure d'Économie, Justus-Liebig-Universität Gießen	Examinatrice
Eric Darmon Professeur d'Économie, Université Paris Nanterre	Examinateur
Ulrich Laitenberger Professeur d'Économie, Télécom Paris, Institut Polytechnique de Paris	Directeur de thèse
Anthony Strittmatter Professeur d'Économie, UniDistance Suisse	Co-encadrant de thèse

Acknowledgements

First and foremost, I would like to thank the many people that have been undispensable when conducting the research and analyses of this dissertation.

My supervisor *Ulrich Laitenberger* has been a constant support and encouragement throughout the journey. He has given me countless advice on several aspects of the projects, starting from which project ideas to push and which subanalyses to look into, to encouraging me to always have the bigger pictures in mind. This dissertation has greatly benefitted from all of our meetings - in person or online during tough Covid times. Also my co-supervisor *Anthony Strittmatter* was always available when I needed advice, which I made use of especially with econometric questions. Both of them always had my best interests in mind and cared not only about the progress of my projects, but also about my workloads and personal life.

I also want to thank my thesis committee for providing valuable inputs before and during my defense. *Grazia Cecere* and *Michael Kummer* have guided me throughout the yearly "mini defenses" and encouraged me to keep up the work. My final committee, consisting of *Marc Bourreau, Grazia Cecere, Fabrice Le Guel, Irene Bertschek, Eric Darmon* and my supervisors has provided excellent ideas, questions, and suggestions, which largerly advanced the final version of this dissertation. I appreciate all the time and efforts that each committee member has put into improving each of the three chapters.

Furthermore, my encouragements during the years have been largely driven by the great co-authors I was lucky to work with. The discussions with *Thomas Niebel, Steffen Viete, Olga Slivko, Michelangelo Rossi* and *Mark Tremblay* always demonstrated that what we do is important and interesting and worth working for. Besides that, meeting and working with each of them was always a lot of fun and marks in the calendar that I looked forward to.

This dissertation has been written within two different institutions: While my everyday work took place at ZEW Mannheim, I enjoyed an even broader network of great scientists by also being part of the economics department of Telecom Paris at IP Paris and CREST. My research stays in Paris and the department workshops in Northern France were great experiences that contributed not only to the quality of my projects, but also to the fun and motivation I had in being part of this community.

At ZEW, I am really greateful for all the wonderful colleagues I shared so many interesting and funny moments with throughout the years. I learned a lot in our seminars and always enjoyed

being part of our easy-going, productive and supportive team. A special thanks belongs to my manager *Irene Bertschek*, who always gave me the freedom to follow my interests and pursue projects and collaborations I enjoyed working with.
Lastly, I am blessed to have a wonderful family. My husband *Arne* is always interested in what I do and also contributed to this thesis by discussing current digital and societal topics with me. Besides that, he was of immensive help in some coding and many practical IT related questions. Also important was the support of my parents and parents-in-law, who took care of our beloved son Levin for countless hours so that I could participate in meetings and continue working when deadlines were approaching.

Thank you to each and every one in this great network of advisers, collaborators, peers, and family. I truly appreciate your support and recognize the contributions you have made to completing this dissertation.

Raphaela Andres

Mühlhausen, June 2024

Contents

List of Figures

List of Tables

Chapter 0

Introduction

0.1 Overview of the Digital Economy

The digitalization has transformed multiple aspects of our lives, such as social interactions, economic transactions and policy making. Over the internet, firms and people can communicate with each other, share data, and sell goods and services. These interactions open new opportunities in the economic and social spheres. Entirely new markets have emerged, such as online market places, social media, and the creator economy. Firms' ability to analyse data has been linked to fostering innovations (Gierten et al. 2021) and to explore avenues to increase firms' production and marketing efficiency (Gal et al. 2019, Borowiecki et al. 2021). Digital platforms allow each individual to participate with very low or zero pecuniary costs in content diffusion and consumption. Examples like online encyclopedias, dating portals, and messenger services showcase the great potential of these platforms to foster a knowledgeable, inclusive, and connected society.

However, the digital economy also creates new challenges. The access to technology is costly in terms of financial and human capital and can constitute a market entry barrier for young and entrant firms. Furthermore, larger firms are more likely to adopt information and communication technologies (ICT, DeStefano, De Backer, and Moussiegt 2017, Calvino and Fontanelli 2023), which can increase the productivity gap between laggard and frontier firms (Gal et al. 2019). Andrews, Criscuolo, and Gal 2016 recognize the slow diffusion of new technologies as one potential reason for the modern productivity paradox.[1]

Furthermore, low marginal costs of participation allow everyone to share hostile opinions and misinformation, which can destabilize democracy and culture, as hate speech on social media harshly demonstrates. According to a survey conducted among the German internet population, almost 60% of the users do not share their political opinion and are less likely to participate in political online discussion due to fear of provoking hateful comments.[2] The growth in openly communicated hatred has been unprecedented and is alarming. Many studies document that hate speech does not solely stay online, but that it impacts individuals, the society, and the economy in the real world. These impacts range from deteriorated well-being of individuals (Allcott,

[1] While new technologies offer great firm-level potentials for productivity as described above, the global productivity growth has experienced a strong decrease in the past 15 years (Andrews, Criscuolo, and Gal 2016). This decrease is of significant magnitude, with growth measures experiencing a cut by half of the pre-slowdown decade (Brynjolfsson, Rock, and Syverson 2019).

[2] Survey "Lauter Hass, Leiser Rückzug": `https://kompetenznetzwerk-hass-im-netz.de/lauter-hass-leiser-rueckzug/`

Braghieri, et al. 2020) to offline hate crimes (Müller and Schwarz 2021, Müller and Schwarz 2023). Economic consequences of online hate speech include the mass withdrawal of advertisers from YouTube in order to mitigate the risk that the advertisers' brand names could be associated with offensive YouTube content.

Policy makers as well as platform managers have recognized the necessity to counteract the above challenges that the digital economy brings along. To increase the diffusion of ICT, most countries fund programs to facilitate the uptake of technology (e.g. the German GRW described in Chapter 1 or the Annual Investment Allowance in the UK[3]. To protect users and advertisers from objectionable content, most social media platforms operating in Europe have joined the *EU Code of conduct on countering illegal hate speech online*[4] and implemented community standards and house rules. However, the platforms' preferred effort level to counteract online hate speech may not fully align with the socially desired level due to high costs of content moderation practices (Madio and Quinn 2023). Also, the platforms may face foregone network effects, since extreme content tends to be spread further than other content, generating more platform participation (Mallipeddi et al. 2021). Therefore, regulations such as the German Network Enforcement Act and the European Digital Services Act (DSA) are intended to further decrease the prevalence of online hate speech on social media platforms. While these regulations do not hold the platforms accountable for all content they are hosting, they oblige large platforms to implement easy reporting mechanisms for users to report illegal contents and to withheld respective contents within a short time frame. One challenge of these regulatory interventions is to reduce illegal content while preserving the freedom of expression.

0.2 Contributions of this Thesis

Although the above-mentioned interventions are of crucial importance to protect the economy and society from the challenges created by the digital economy, empirical evidence on the effectiveness and potential side effects of these policy changes and regulations is scarce. Especially the issue of far reaching online hate speech, a relatively new phenomenon, has to be tackled with new regulatory designs which need to be analysed in depth.

[3] `https://www.gov.uk/capital-allowances/annual-investment-allowance`

[4] `https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en`

This thesis empirically investigates three of these interventions to shed light on the new environment which policy makers and managers face when designing rules and regulation in the digital economy. It demonstrates that the implementation as well as the impact of policy changes and regulation can differ in the digital world compared to the non-digital world. Therefore, each of the chapters of this doctoral thesis points at a special characteristic of the digital world with regard to policymaking that differs from the offline world. These findings are crucial in order to design well targeted policies and avoid unintended side effects and are of interest to policy makers, managers, and the general public.

The first chapter of this doctoral thesis, co-authored with Thomas Niebel and Steffen Viete and published in *Telecommunications Policy*, uncovers an unintended policy effect of an investment funding program, which emerges since the policy has not been updated to consider technological advances. This chapter was featured in the media outlets *Wirtschaftswoche*[5] and *heise*[6]. The remainder of the thesis exploits large data on user-generated content on digital platforms. The second chapter, co-authored with Olga Slivko, documents the causal effectiviness of the first regulation tackling the diffusion of online hate speech. It also shows that other posting styles, such as posting frequencies and posting of images, are not altered by the regulation. This chapter was awarded the "Best Student Paper" of the International Telecommunications Society Conference in 2021 and was taken up by several media outlets such as *Tagesspiegel Background*[7] and *Deutsche Welle*[8]. The third chapter, co-authored with Michelangelo Rossi and Mark Tremblay, reveals cross-platform impacts of a rule change on one platform, emphasizing that regulatory interventions on one platform cannot be considered in isolation.
All of these chapters are publicly available.[9]

[5] `https://www.wiwo.de/technologie/digitale-welt/digitale-transformation-der-staat-pumpt-milliarden-in-digitalisierung-und-bremst-damit-cloud-computing-aus/26153976.html`
[6] `https://www.heise.de/news/Foerderprogramme-bremsen-Cloud-Investitionen-4885144.html`
[7] `https://background.tagesspiegel.de/digitalisierung/netzdg-zeigt-wirkung-auf-twitter?`
[8] `https://www.dw.com/en/twitter-how-elon-musk-changed-x-in-one-year/a-67220997`
[9] Chapter 1:
`https://www.sciencedirect.com/journal/telecommunications-policy/vol/48/issue/1`
Chapter 2:
`https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4013662` (slightly earlier version)
Chapter 3: `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4672028`

Investment Subsidies and the Diffusion of Digital Services

The access to technology is costly and can constitute a market entry barrier that could lessen the benefits of competition. To facilitate the investment in ICT for firms in economically lagging regions, the German government offers an investment subsidy, the "Joint task for the improvement of the regional economic structure". However, targeting digitization investments to facilitate large upfront investments has been necessary *before* the possibility of adopting digital services like cloud computing, which constitute ongoing expenses. In fact, this lagging-behind of the policy with regard to rapidly changing technologies might even hinder the development of the targeted firms, since cloud service are associated with various benefits for the firm, such as increased flexibility (Jin and McElheran 2017) and freeing up resources for innovation and marketing (Columbus 2013).

This chapter of the doctoral thesis investigates the hypothesis that public investments schemes can reduce the uptake of cloud computing services of firms. This hypothesis can be explained by the following example: A firm needs to increase its data storage capacities. It now has the options to purchase a new server or to purchase the storage capacity via a cloud service provider. If there is an investment policy in place which lowers the cost for investing in a new server, the firm might be incentivized to make use of the funding and purchase the new server. Notably, due to the scalability and possibility to constantly update to newest technologies in the cloud, the investment in the new server might not be the optimal choice for the firm in the long-run — even with the investment funding. Furthermore, since cloud computing is also associated with long-run benefits on the economy level (such as lowering market entry barriers, creating new jobs and increasing competition (OECD 2014)), and the funding costs of the investment subsidy, the subsidy might also be a non-optimal macroeconomic and political choice.
To investigate the hypothesis, my coauthors and I exploit how variation in the eligibility as well as the size of potential investment subsidies of German firms affects the firms' propensity to adopt cloud services. Our results are in line with the above hypothesis of an unintended policy effect: Firms which are eligible for investment funding have a lower probability to use cloud computing services. Furthermore, the higher the potential subsidy for investments, the lower the incentive for firms to adopt cloud services. However, due to the lack of suitable data, these relationships should not be interpreted as causal, but as conditioned correlations. The results hold after conditioning on various factors such as the firms' size (employment and turnover),

location, sector, age, and digitization level.

These results are of great interest to policy makers that are involved in designing policies fostering the digital transformation. It should be reconsidered if current incentive schemes for the digital transformation of the economy are fit to the current needs and possibilities brought about by the fastly changing environment of the digital economy.

The First Regulation to Reduce Online Hate Speech

Online hate speech has tremendous impacts on individual well-being (Allcott, Braghieri, et al. 2020), political engagement (Enikolopov, Makarin, and Petrova 2020), and social cohesion (Bursztyn et al. 2019). Due to the novelty of the threats and challenges brought by digital platforms, new regulations are needed to tackle these issues. The second chapter of this doctoral thesis analyses the effectiveness of the first regulation that aims at restraining hate speech on large social media networks, the German Network Enforcement Act (NetzDG). This law obliges large social media platforms in Germany to implement simple procedures for users to report hateful content and requires the social networks to remove hateful content within a short time frame. We exploit the implementation of the NetzDG in a quasi-experimental approach and measure the causal impact of the law on the prevalence and volume of hateful content in a German-speaking target group on the platform Twitter (now X). For measuring hate speech, my coauthor and I use pre-trained algorithms, which have demonstrated quite accurate performance (Mondal, Silva, and Benevenuto 2017, ElSherief et al. 2018, Han and Tsvetkov 2020). We find evidence of a significant and robust decrease in the intensity and volume of hate speech in tweets tackling sensitive migration and religion-related topics. Importantly, tweets addressing other topics as well as the tweeting style of users are not affected by the regulation, which is in line with its aim. These findings are very robust and contribute to a sound understanding of the effectiveness and the functioning of the law, which is crucial to make informed decisions on the issue of hate speech regulation. This becomes even more important in light of the fact that the respective regulation – the German NetzDG - was used as a blueprint for similar laws in other countries (Tworek and Leerssen 2019), including the recently implemented European Digital Services Act.

Cross-Platform Effects of Interventions

Digital platforms in the Creator Economy have thrived on a multi-sided business model, capitalizing on the cross-network effects of three primary stakeholders: content creators (CC), content consumers (users), and advertisers. However, matching the needs of different stakeholders is a complex challenge, as evidenced by the impact of the YouTube "Adpocalypse" in 2017, when major advertisers fled YouTube due to concerns about their advertisements appearing alongside objectionable content. At the same time, CCs face very low or zero costs for setting up a profile on these platforms in the Creator Economy, providing them with very low barriers to entry and to allow them to switch between or establish a presence on multiple platforms simultaneously (multi-homing). Hence, when a platform aims to explore changes in its business model, it is critical for platforms to analyze the response of all stakeholders. This includes their potential migration to another platform or increased efforts in creating content on competing platforms.
In this chapter of the doctoral thesis, my coauthors and I offer theoretical and empirical evidence of how CCs and users react to changes in a platform's moderation policy. Our focus centers on two of the most successful platforms for CCs: YouTube and Patreon. YouTube primarily monetizes content through advertising revenue, while Patreon is a platform designed to help CCs generate income directly from their supporters through monthly subscriptions. Our analyses explore the responses by CCs that use both Youtube and Patreon to YouTube's content moderation policies following the "Adpocalypse". We find that these CCs shift their efforts toward the subscription fee model instead of the ad-based model; as a result, users subsequently increase their use of Patreon through memberships, comments, and likes. However, we also find that Youtube's content moderation along with the shift by CCs and consumers that follows results in an increase in overall toxicity on Patreon.
As only a few empirical studies examine cases of competition between asymmetric platforms, these results are relevant to several stake holders. For CCs they imply that being open to different business models and platforms can protect their monetization from platform-level shocks. For platform managers, the results highlight the importance to account for substitutive behaviour of the CCs in order to evaluate potential impacts of rule changes. From a policy perspective, this chapter suggests that regulating only large online companies, as it is the case in current legislation, may not fully achieve its goals due to network effects and multi-homing options, as the CCs and users can can simply migrate to another platform.

Chapter 1

Do Capital Incentive Policies Support Today's Digitization Needs?

Published as

Andres, Raphaela, Thomas Niebel and Steffen Viete (2024), Do Capital Incentive Policies Support Today's Digitization Needs?, *Telecommunications Policy* 48(1).

1.1 Introduction

Cloud computing services are associated with various potential benefits on the firm as well as the aggregate economy level. As such, adopting cloud computing services reduces necessary upfront investments in IT infrastructure for firms and consequently lowers the market entry barriers, which in turn can lead to increased market competition, innovation, and employment (OECD 2015). Through the commonly used "pay-as-you-go" payment method of cloud services, firms can rapidly adopt cutting edge technology that fits their needs in terms of functionalities and scale (OECD 2014), which has shown to be even more beneficial in insecure times like the Covid pandemic. Furthermore, cloud computing services facilitate the working with and sharing of large data pools, commonly known as big data analysis. This allows analysing costumers' needs to improve products and services (Gierten et al. 2021) and has been shown to increase overall firm productivity (Borowiecki et al. 2021).

Due to the above benefits of cloud computing on the firm and on the aggregate level, firms' adoption of cloud services is increasing rapidly. Public cloud spending in Europe is predicted to experience a fivefold increase between 2016 and 2023 to about 100 billion Euro yearly (Statista 2022). Nevertheless, it should be of a policy maker's interest to foster the adoption of cloud services of firms. However, digitization policies are still mostly targeting digitization investments, which has been necessary *before* the possibility of adopting digital services. The targeting of digitization investments might not only be a missed opportunity to increase the usage of digital services in the economy, but might even lower the uptake of digital services. This hypothesis can be explained by the following example: A firm needs to increase its data storage capacities. It now has the options to purchase a new server or to purchase the storage capacity via a cloud service provider. If there is an investment policy in place which lowers the cost for investing in a new server, the firm might be incentivized to make use of the funding and purchase the new server. Notably, due to the scalability and possibility to constantly update to newest technologies in the cloud, the investment in the new server might not be the optimal choice for the firm in the long-run - even with the investment funding. Furthermore, due to the long-run benefits of cloud computing on the aggregate economy level described above and the funding costs of the investment subsidy, the subsidy might also be a non-optimal political choice.

This paper investigates the hypothesis that public investments schemes can reduce the uptake of cloud computing services of firms, which would be an unintended policy effect. To do so, we

exploit variation in the eligibility as well as the size of potential investment subsidies of German firms to analyse the relationship with the firms' propensity to adopt cloud services.
Our results are in line with the above hypothesis of an unintended policy effect: Firms which are eligible for investment funding have a lower probability to use cloud computing services. Furthermore, the higher the potential subsidy for investments, the lower the incentive for firms to adopt cloud services. Due to the lack of suitable data, these relationships should not be interpreted as causal, but as controlled correlations. Yet, the results hold after controlling for various factors such as the firms' size (employment and turnover), location, sector, age, and digitization level.

1.2 Related Literature

The basic characteristics and the resulting potential benefits of cloud computing services have been known for some time now (see e.g. Armbrust et al. 2010; Mell and Grance 2011; OECD 2014). One key driver for the increasing demand for public cloud services are their rapidly falling prices (see e.g. Coyle and Nguyen 2018; Byrne, Corrado, and Sichel 2021). However, empirical analyses investigating the causal impact of cloud computing on firm performance are still surprisingly scarce. This is mostly driven by the fact that high quality firm-level data on cloud computing adoption was not readily available (Haug, Kretschmer, and Strobel 2016).[10] One exemption is the paper by DeStefano, Kneller, and Timmis 2020. Based on UK firm-level data, they find that cloud computing is primarily beneficial for younger firms. Younger firms adopting cloud computing services see an increase in revenues, employment and productivity. Similar, Jin and McElheran 2017 find that younger firms have higher gains from IT-services (their measure of cloud computing) than older firms. Duso and Schiersch 2022 use administrative firm-level data for Germany and quite surprisingly find *no* evidence that the use of public cloud services is a substitute for own IT-investments. However, they do find a positive and causal relationship between cloud computing adoption and labour productivity at least in the manufacturing and information and communication services sector. Gal et al. 2019, combining firm and industry-level data, also find a positive association between cloud adoption and productivity, but again only for the manufacturing sector. Apart from these papers measuring the direct effects of cloud computing on firm performance, there is some (mostly descriptive) evidence that cloud

[10] This has improved in recent years. However, firm surveys still often only measure whether or not firms use cloud technologies and do not measure the intensity of use (Ker 2021).

computing is related or even a prerequisite for the adoption of more advanced IT (see e.g. Zolas et al. 2020; Gierten et al. 2021; Cho et al. 2023). With respect to the adoption of ICT in general (Haller and Siedschlag 2011; DeStefano, De Backer, and Moussiegt 2017) and the adoption of more advanced technologies like artificial intelligence (Calvino, Samek, et al. 2022; Calvino and Fontanelli 2023), larger firms are more likely to adopt these technologies.

Additionally to the literature on cloud computing and digital technology adoption in general and their respective impact on firm performance, our paper is also related to the impact of investment subsidies on economic outcomes. Maffini, Xing, and Devereux 2019 for example find an economically significant increase in the investment rates of UK firms for those that are eligible for accelerated depreciation allowances on investment. The paper by Criscuolo et al. 2019 finds that a 10-percentage point increase in the maximum investment subsidy leads to a 10 percent employment increase in the manufacturing sector in the UK. Both papers show that firms adjust their behaviour when there is public funding. There is also a number of papers investigating the general effects of the German GRW subsidy[11], which is the policy scheme we exploit in our empirical setting. Alecke and Mitze 2023 show that changes in maximum funding are related to an increase in the average investment intensity as well as newly created jobs. Dettmann, Titze, and Weyh 2023 also find positive employment effects of GRW grants but point out that there is effect heterogeneity based on firm characteristics.[12] The paper by Siegloch, Wehrhöfer, and Etzel 2022 comes to a similar conclusion with respect to GRW subsidies on employment. A decrease in the subsidy rate leads to reduced employment in the manufacturing sector.

1.3 Data

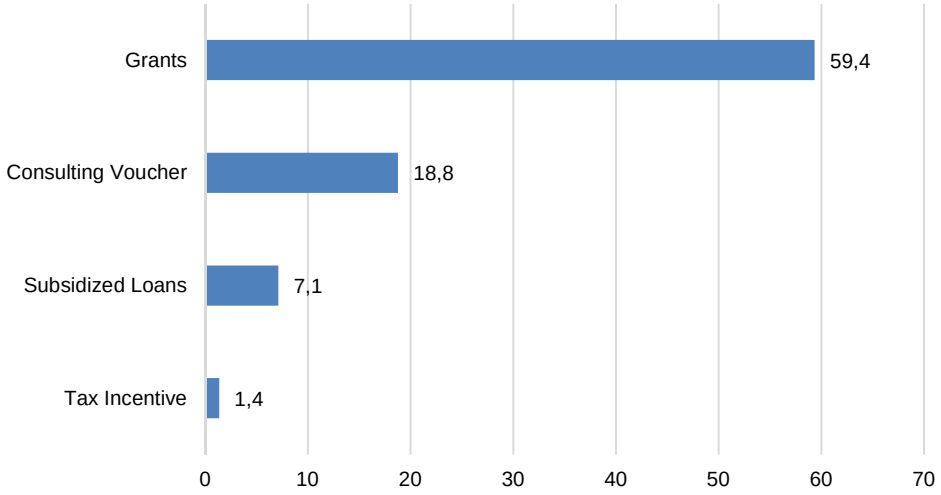
1.3.1 The German GRW Subsidy

Traditionally, public financial support in Germany is directed towards rewarding investments, often times through grants or loans. As shown in Figure 1.1, grants are the most important policy incentive for digitization projects in the German information sector: Of the firms that applied for investment support, 60% indicated that they applied for receiving an investment grant. Other

[11] For a detailed description of the German GRW funding scheme see Section 1.3.1.
[12] We therefore include a large number of control variables. See Table 1.3.

policy incentives, such as consulting vouchers, subsidized loans or tax incentives, are by far less prevalent.

Figure 1.1: Relevance of Different Policy Incentives for Digitization Projects in the German Information Sector



Note: Share of firms among those which applied for support for digitization projects.

Source: ZEW Economic survey of the information sector (2019).

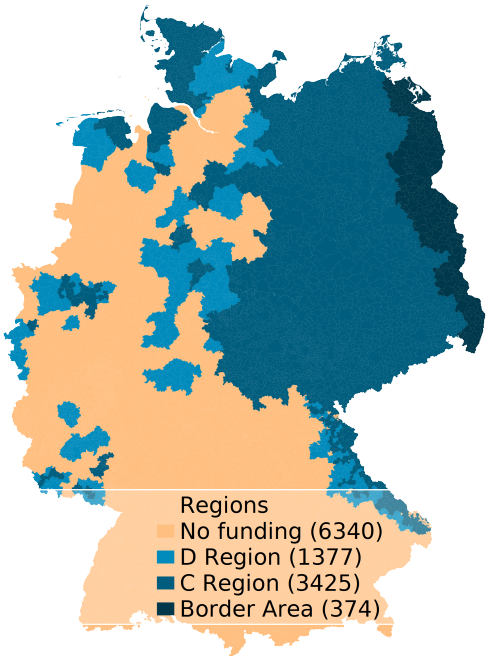
Therefore, we chose to investigate the relationship between the access to the largest German investment funding that is based on providing investment grants, which is called "Joint task for the improvement of the regional economic structure" (GRW). One of the project's deliberate goals is to support private businesses in economically lagging regions through funds for physical capital investment projects for expansion and diversification of production or for fundamental changes to the production process. A second objective of GRW is the support of public infrastructure, which made up 30% of all grants between 1995 and 2014 (Deutscher Bundestag 2016).[13] Targeted regions eligible for funding and the maximum shares of the investment costs which can be funded were newly defined in 2014. Eligible regions are chosen according to an evaluation of various indicators based on unemployment, gross salaries, expected employment, and infrastructure. The German Federal States are responsible for implementing GRW, i.e. they decide about the allocation of funds to eligible projects (Deutscher Bundestag 2014). Funding is available for

[13] We note that infrastructure funds in the GRW framework should either be neutral towards the firms' decision between investments in ICT assets versus acquisitions of ICT services, or, in case they are used for broadband infrastructure, they should indirectly incentivize cloud adoption. This would downward bias potential negative effects of investment incentives on cloud adoption.

specific investment projects and eligible costs are capital expenditures or personnel costs.[14] Maximum funding rates of the investment costs vary regionally and by firm size. To receive any GRW funding, the investments can, but do not need to be digitization specific, but should aim at keeping or increasing employment within the firm.[15]

Funds are available in the whole of Eastern Germany and, with lower funding rates, in various regions in West Germany. Maximum funding intensities, i.e. the shares of the total investment costs which can get funded, were assigned based on the region's previous economic output (Figure 1.2). The regional variation in eligibility for public funding within the scope of the GRW at the municipal level will be used in this paper to investigate the relation between public investment incentives and firms' use of cloud computing and other IT-assets.

Figure 1.2: GRW 2014 Regions



Note: The figure shows the distribution of the different funding areas as described in Table 1.1. The numbers in brackets correspond to the number of individual regions within the coloured funding areas.

Source: Authors' illustration based on BAFA 2019.

[14] Figure 1.3 in the Appendix provides information on the development of GRW cases and funding over time.

[15] Some industries (e.g. retail trade, energy and water, construction and transport) are exempted from the eligibility for GRW funding. Our analyses acknowledge this by dropping the firms in the respective industries.

Figure 1.2 plots the GRW regional aid map, which came into effect in mid-2014. While location determines whether a firm has access to GRW funding, the map additionally illustrates variation in the maximum funding rates across regions and by firm's SME status. Whereas the whole of Eastern Germany has access to GRW funding, in Western Germany only selected regions are addressed by GRW. The highest funding rates apply to regions in Eastern Germany, which are located at the border to Poland. Since GRW funding is targeted towards economically weaker regions and is implemented by the federal states, we need to take account for confounding regional characteristics in our empirical analysis. We will therefore control for regional states, as well as the municipalities' population density and broadband quality at the firm level, which both proxy for regional economic performance at the most granular level.

Table 1.1 displays the maximum funding rates for the funding period 2014-2020 as determined by the GRW region and SME status.[16] For instance, a small firm located in a "D region" can apply for a grant that amounts to 20% of the investment costs of the respective project. Funding rates are higher for small enterprises in each region. Maximum funding rates range up to 40 % of the eligible investment costs.

Table 1.1: Maximum Incentive Rates in the GRW Programme

Region	Small enterprise	Medium enterprise	Large enterprise
Border area	40 %	30 %	20 %
C region	30 %	20 %	10 %
D region	20 %	10 %	200.000€

Notes: Percentage of eligible investment costs. The lowest maximum available funding rates in D regions are 20% for small- and 10% for medium sized enterprises. For large enterprises this limit is set in an absolute value, 200.000€.
Source: Deutscher Bundestag 2014.

In addition, the GRW also serves as a coordination framework for other policies in Germany, which aim at supporting regional development. Thus, the same regions are addressed by the European Regional Development Fund (ERDF), as well as the ERP Regional Promotion Programme by the German government-owned development bank (KfW). We note that these two policies also target investments, either through grants (ERDF) or through loans (ERP). Therefore, when we simply explore regional variation in access to investment incentives, we capture these

[16] See the EU recommendation 2003/361 (`http://data.europa.eu/eli/reco/2003/361/oj`). In particular, we will treat firms with less than 50 employees and annual sales up to 10.000€ as small, firms with less than 250 employees and sales up to 50.000€ as medium, and firms beyond as large firms.

policies along with GRW. In contrast, exploiting variation in the maximum funding rates is specific to the GRW programme. We also note that other incentives for digitization projects, such as consulting vouchers, typically do not overlap with the regions defined by GRW.

1.3.2 Final Data Set

The econometric analyses rely on a data set which combines information from various administrative sources. This includes administrative data for cloud adoption, which stem from the *Survey on ICT Usage and E-Commerce in Enterprises* administered by Eurostat. In addition, we use the German administrative business registry that contains additional information on firm characteristics. Finally, we rely on policy data, which provides specifics about the context and eligibility of the GRW grant scheme.

The primary data source is administrative data on the use of cloud computing by firms. Under the administration of Eurostat, information on cloud computing and other ICT variables are collected by means of a business survey by each country annually through their office of national statistics, thus resulting in reasonable consistency in terms of questions asked and technologies covered across countries over time. The German data set provided by the German Federal Statistical Office (destatis) is called "Erhebung zur Nutzung von Informations- und Kommunikationstechnologien in Unternehmen" (henceforth ICT survey).[17] Aside from Schivardi and Schmitz 2020 and Duso and Schiersch 2022, this paper is among the first to exploit this data set for firm-level analyses. Information on cloud adoption pertains to the years 2014 and 2016.

In order to locate firms in municipalities, we match the administrative ICT survey with the German business registry (Unternehmensregister), which, in addition to regional identifiers, contains information on the firms' industry affiliation, sales, number of employees, and firm age.

Data on the GRW programme has been acquired through the German Federal Office of Economic Affairs and Export Control (BAFA) as well as the German Federal Ministry for Economic Affairs and Energy (BMWi). The data contain information at the municipal level on

[17] Source: RDC of the Federal Statistical Offices and Offices of the Länder, "Erhebung zur Nutzung von Informations- und Kommunikationstechnologien in Unternehmen", survey years 2014 and 2016, own calculations (`www.doi.org/10.21242/52911.2014.00.00.1.1.0` and `www.doi.org/10.21242/52911.2016.00.00.1.1.0`).

whether or not a municipality is eligible to GRW grants, maximum funding rates, and approved funding for the years 2000 until present.

Table 1.2 shows descriptive statistics of the pooled cross section that we will rely on for our analyses.[18] Our sample comprises 8,540 observations on cloud computing usage throughout Germany, out of which 4,670 firms are observed in 2014 and 3,870 firms are observed in 2016. Cloud computing is used by 21% of the firms in the sample. Within the average observation, 55.1% of the employees have access to the internet and 18.1% of the employees are equipped with a mobile internet connection. 34% of the firms were eligibility for GRW funding and 50% of the firms employed own IT-staff.

Table 1.2: Summary Statistics of the Estimation Sample

	N	Mean	Median	SD
Cloud computing	8540	0.21	0	0.41
Number of employees	8540	423.7	75	4434.5
Sales	8540	135478.4	10044.2	2288269.5
% of employees with internet connection	8540	55.1	50	33.3
% of employees with mobile internet connection	8540	18.1	10	22.6
Eligibility for GRW funding	8540	0.34	0	0.47
Employment of own IT-staff	8540	0.50	1	0.50

Source: BAFA 2019 and Eurostat ICT Survey.

1.4 Estimation Strategy

To investigate whether the investment scheme discourages the use of cloud computing services, we exploit regional variation in the access to GRW funding as well as the differing incentive rates as described in Table 1.1.

Unfortunately, the survey is constructed as a rotating panel, such that the overlap of firms participating in both years is not large enough to run robust panel analyses. Hence, we will run pooled cross sections estimations and restrain from claiming causal relations, as our results should be interpreted as controlled correlations. For the sample of firms described in Section 1.3, we know the exact location, firm size, and usages of technologies including cloud computing and several potential control variables. Hence, we can match these firms to the respective maximum

[18] See Table 1.8 in the Appendix for summary statistics of all variables.

incentive rate and an indicator variable for general eligibility for GRW funding. However, we do not observe if a firm in our sample indeed received GRW funding. Therefore, the following analyses estimate the intention to treat (ITT) effect of the GRW funding on the propensity to use cloud computing services by estimating logit regressions on the firm level.

Equation 1.4.1 estimates the relation between the access to GRW funding and the propensity to use cloud computing services and hence estimates the extensive margin of the funding effect.

$$\mathrm{CC}_i = \beta_0 + \beta_1 \mathrm{Treat}_i + \beta_2 X_i^{'} + \varepsilon_i \tag{1.4.1}$$

CC_i is an indicator which equals one if firm i uses cloud computing and zero otherwise. The vector $X_i^{'}$ represents a number of control variables such as the logarithm of a firm's sales and the broadband speed availability. Broadly speaking, we control for the organizational size, industry, the firms' general ICT intensity, and regional characteristics (see Table 1.3 for the full list of control variables). The parameter of interest is β_1, which measures the change in the propensity of a firm to use cloud computing services that is due to being eligible for GRW funding, given the location, industry, year, and further control variables as indicated in Table 1.3. ε_i is the idiosyncratic error term that captures unobservable characteristics related to the firms' propensity to adopt cloud services.

Instead of the binary information if a firm is eligible for GRW funding or not, Equation 1.4.2 exploits the variation in maximum funding rates firms can receive depending on their location and size.

$$\mathrm{CC}_i = \beta_0 + \beta_1 \mathrm{Grant}_i + \beta_2 \mathrm{Grant}_i^2 + \beta_3 X_i^{'} + \varepsilon_i \tag{1.4.2}$$

Hence, this analysis investigates the intensive margin of being eligible for funding between 0% and 40% of physical capital investments and incorporates this information in the firm-specific variable $Grant_i$. We also include the maximum incentive rate in squared terms in order to allow for a more flexible relation between the funding rates and the propensity to adopt cloud services. As such, the funding rates could have a decreasing marginal effect on the likelihood to use cloud computing services.

1.5 Empirical Results

The following section presents the results of the econometric analyses of the relation between access to the regionally targeted investment grants as determined by the GRW at the extensive and intensive margin. The analyses first assess the relation between cloud use and eligibility for investment grants as well as the maximum funding rates the firm is able to apply for.

Table 1.3 presents the results for the extensive margin of being eligible for public investment grants. As the dependent variable is the binary outcome if firm i uses cloud services or not, we estimate Equation (1.4.1) using logit models and present the average marginal effects in Table 1.3. In Column (1) we estimate a parsimonious model in which we only include a full set of dummy variables for federal states, industry, and year. In Columns (2) and (3) we additionally include into our model the log number of employees, as a control for firm size, as well as the log of sales. This considerably reduces the measured relation between treatment status and cloud adoption. In addition, in Column (4) we control for the firm's use of internet based ICT by including the share of employees with access to the internet in general and with access to the mobile internet. Furthermore, we include the firm's fixed-line internet quality by adding a set of dummy variables denoting internet speed and account for the firm's age in logarithmic terms as well as the population density in the municipality. The last column additionally includes dummies for being a medium or large enterprise in order to further isolate the impact of the public funding rather than being a SME.

In all specifications, we find a negative and statistically significant relation between access to public investment incentives and the propensity to use cloud computing. Looking at Column (4) as our preferred specification, we find that having access to investment incentives decreases the propensity to use cloud computing by 2.5 percentage points. This effect is statistically significant at the 10% level.

Looking at other variables in the model, we find that firm size is an important determinant of cloud adoption. According to the estimates in Column (4), a one percent increase in the number of employees is associated with a 4.6 percentage point increase in the propensity to adopt cloud. Moreover, internet access is an important determinant for the use of cloud technologies. Looking at Column (4) again, a one percent increase in the share of employees with access to the internet relates to a 0.2 percentage point increase in the propensity to adopt cloud. Beyond

the general use of internet in the firm, a respective increase in the share of employees with access to mobile internet technologies increases the likelihood to adopt cloud by 0.1 percentage points. Furthermore, our estimation results underline the importance of internet quality for the use of cloud technologies. We find statistically significant and positive effects for the indicators denoting internet access with 2 Mbit/s and beyond. Interestingly, the effects get larger for higher bandwidth up to 30 Mbit/s while there is no increase in the effect when moving further to 100 Mbit/s. Overall, these results suggest that there is a decreasing return to internet speed in terms of firms' cloud adoption. In interpreting these results one has to keep in mind that the data refer to the years 2014 and 2016. Finally, cloud adoption is more likely in younger firms as denoted by the negative marginal effect of firm age on cloud adoption.

The same picture emerges when looking at the intensive margin of the GRW funding. In addition to the firms' location, maximum GRW funding rates are also determined by the firms' SME status. Again, note that the treatment dummy from our first set of results in Table 1.3 captures access to a multitude of regionally targeted public investment incentives besides the GRW, such as the ERDF. In contrast, the maximum funding rates only refer to the GRW programme. Table 1.4 shows that eligibility for a higher grant (i.e., receiving a higher share of the capital investment as a non-repayable monetary grant) reduces the propensity to use cloud computing services. Even after controlling for several aspects such as firm location, industry, and digitization level, the relationship stays significantly negative (Table 1.4).

Table 1.3: Cloud Computing and Access to Regional Incentives - Logit Regression - Average Marginal Effects

	(1)	(2)	(3)	(4)	(5)
Treat	-0.047***	-0.036***	-0.033**	-0.025*	-0.024*
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
ln(Employees)		0.058***	0.033***	0.046***	0.047***
		(0.003)	(0.006)	(0.005)	(0.006)
ln(Sales)			0.025***	0.007	0.007
			(0.005)	(0.004)	(0.004)
% of employees with internet connection				0.002***	0.002***
				(0.000)	(0.000)
% of employees with mobile internet connection				0.001***	0.001***
				(0.000)	(0.000)
Below 2Mbit/s				0.032	0.032
				(0.033)	(0.033)
Between 2 Mbit/s and 10 Mbit/s				0.090***	0.091***
				(0.025)	(0.025)
Between 10 Mbit/s and 30 Mbit/s				0.109***	0.109***
				(0.025)	(0.025)
Between 30 Mbit/s and 100 Mbit/s				0.138***	0.138***
				(0.026)	(0.025)
More than 100 Mbit/s				0.110***	0.110***
				(0.026)	(0.026)
ln(Age)				-0.016***	-0.016***
				(0.006)	(0.006)
Population density				0.007	0.006
				(0.006)	(0.006)
Medium					-0.011
					(0.014)
Large					-0.009
					(0.022)
Industry Effects	Yes	Yes	Yes	Yes	Yes
Fed. State Effects	Yes	Yes	Yes	Yes	Yes
Year Effects	Yes	Yes	Yes	Yes	Yes
Pseudo R^2	0.036	0.077	0.081	0.119	0.119
Observations	8540	8540	8540	8540	8540
Log likelihood	-4203.210	-4020.500	-4003.134	-3839.402	-3838.984

Notes: Robust standard errors in parentheses, * p<0.10, ** p<0.05, *** p<0.01.
All models include an intercept.

Table 1.4: Cloud Computing and Incentive Rates - Logit Regression - Average Marginal Effects

	(1)	(2)	(3)	(4)	(5)
Grant	-0.330***	-0.320***	-0.297**	-0.240**	-0.230**
	(0.117)	(0.117)	(0.117)	(0.116)	(0.116)
ln(Employees)		0.058***	0.033***	0.046***	0.047***
		(0.003)	(0.006)	(0.005)	(0.006)
ln(Sales)			0.025***	0.007	0.007
			(0.005)	(0.004)	(0.004)
% of employees with internet connection				0.002***	0.002***
				(0.000)	(0.000)
% of employees with mobile internet connection				0.001***	0.001***
				(0.000)	(0.000)
Below 2Mbit/s				0.032	0.031
				(0.033)	(0.033)
Between 2 Mbit/s and 10 Mbit/s				0.090***	0.091***
				(0.025)	(0.025)
Between 10 Mbit/s and 30 Mbit/s				0.109***	0.109***
				(0.025)	(0.025)
Between 30 Mbit/s and 100 Mbit/s				0.138***	0.138***
				(0.026)	(0.025)
More than 100 Mbit/s				0.110***	0.110***
				(0.026)	(0.026)
ln(Age)				-0.017***	-0.016***
				(0.006)	(0.006)
Population density				0.007	0.007
				(0.006)	(0.006)
Medium					-0.009
					(0.015)
Large					-0.008
					(0.023)
Industry Effects	Yes	Yes	Yes	Yes	Yes
Fed. State Effects	Yes	Yes	Yes	Yes	Yes
Year Effects	Yes	Yes	Yes	Yes	Yes
Pseudo R^2	0.041	0.077	0.081	0.119	0.119
Observations	8540	8540	8540	8540	8540
Log likelihood	-4179.222	-4020.400	-4002.856	-3838.925	-3838.700

Notes: Robust standard errors in parentheses, * p<0.10, ** p<0.05, *** p<0.01.
All models include an intercept.

1.6 Discussion

The econometric results presented in Section 1.5 suggest that the public funding scheme GRW discourages firms to adopt cloud computing services. Since cloud computing services are associated with various benefits at the firm and the total economy level, this is a prime example for an unintended policy effect. Due to the low number of firms in the data and the low overlap of firms in the two survey years, our estimation strategy is restricted to showing controlled correlations which should not be interpreted as causal. However, the hypothesis of the unintended policy effect of disincentivizing cloud usage is strengthened by the fact that we do not find significant correlations between the eligibility of the GRW funding and the employment of IT-staff (see Table 1.5).

Table 1.5: Employment of Own IT-Staff and Access to Regional Incentives - Logit Regression - Average Marginal Effects

	(1)	(2)	(3)	(4)	(5)
Treat	-0.048***	-0.016	-0.010	-0.001	-0.001
	(0.016)	(0.014)	(0.013)	(0.013)	(0.013)
ln(Employees)		0.186***	0.114***	0.143***	0.151***
		(0.003)	(0.008)	(0.006)	(0.008)
ln(Sales)			0.068***	0.034***	0.037***
			(0.006)	(0.005)	(0.006)
Additional Controls	No	No	No	Yes	Yes
Additional Controls + Size Dummies	No	No	No	No	Yes
Industry Effects	Yes	Yes	Yes	Yes	Yes
Fed. State Effects	Yes	Yes	Yes	Yes	Yes
Year Effects	Yes	Yes	Yes	Yes	Yes
Pseudo R^2	0.066	0.277	0.296	0.363	0.363
Observations	8540	8540	8540	8540	8540
Log likelihood	-5528.700	-4279.171	-4167.838	-3770.161	-3769.068

Notes: Robust standard errors in parentheses, * p<0.10, ** p<0.05, *** p<0.01.
All models include an intercept. For the list of additional control variables see Table 1.3.

Furthermore, we analyse if the GRW funding has heterogeneous impacts on the adoption of basic and more complex cloud services. We refer to the classification into basic and complex cloud services by Eurostat, which is shown in Table 1.6.

Table 1.6: Definition of Cloud Services by Degree of Complexity

Use of cloud computing service	**Basic cloud**	**Complex cloud**
Email Office Software Storage of Files	At least one	At least one
Hosting the Enterprise's database(s) Finance Software CRM Processing	None	At least one

Source: Eurostat 2018.

Interestingly, we only find significant relations between the intensive and extensive margin of the GRW eligibility and the propensity to adopt basic cloud services. While the average marginal effects of the funding to use basic cloud services stays significantly negative after controlling for several factors such as industry, firm size, and digitization level, we find no such relation to the usage of complex cloud services (see Table 1.7). This finding is especially relevant on the aggregate level, as one major economic benefit of cloud services is the reduction of market entry barriers for new firms, which could increase competition, employment, and innovativeness of markets (OECD 2015). The finding that especially the adoption of basic cloud services is reduced suggests that this economically relevant benefit of cloud computing comes less into play due to the public funding scheme.

Table 1.7: Cloud Computing by Complexity and Access to Regional Incentives - Logit Regression - Average Marginal Effects

	Basic cloud		Complex cloud	
	(1)	(2)	(3)	(4)
Treat	-0.017**	-0.017**	-0.006	-0.006
	(0.008)	(0.008)	(0.011)	(0.011)
ln(Employees)	0.008**	0.006*	0.021***	0.022***
	(0.003)	(0.004)	(0.004)	(0.005)
ln(Sales)	0.003	0.002	0.003	0.004
	(0.003)	(0.003)	(0.003)	(0.003)
Additional Controls	Yes	Yes	Yes	Yes
Additional Controls + Size Dummies	No	Yes	No	Yes
Industry Effects	Yes	Yes	Yes	Yes
Fed. State Effects	Yes	Yes	Yes	Yes
Year Effects	Yes	Yes	Yes	Yes
Pseudo R^2	0.042	0.043	0.119	0.120
Observations	8540	8540	8540	8540
Log likelihood	-1959.383	-1958.454	-2481.183	-2480.832

Notes: Robust standard errors in parentheses, * p<0.10, ** p<0.05, *** p<0.01.
All models include an intercept. For the list of additional control variables see Table 1.3.

1.7 Conclusion

Our empirical results are in line with the hypothesis that policies which foster firm investments discourage the adoption of cloud computing services. This finding is due to the perceived substitutability of investing in own IT infrastructure and purchasing digital services. Although in the long-run, it might have been optimal for firms to purchase digital services as this allows to flexibly adjust their digitization needs and costs, the lowered prices of IT investments due to the public funding incentivizes the IT investment. This main result should be thought-provoking, as the possibility to adopt cloud services promises large investment savings for the firm as well as the government, as the funding scheme should be ceased due to its unintended effect on the uptake of digital services.

We consider this finding an unintended policy effect, since the adoption of cloud services has been linked to various benefits on the firm as well as on the aggregate economy level. Therefore, our results are of great interest to policy makers that are involved in designing policies

fostering the digital transformation. It should be reconsidered if incentive schemes for the digital transformation of the economy should be adapted to today's needs and possibilities by being broadened to also incentivize the uptake of cloud based digital services.

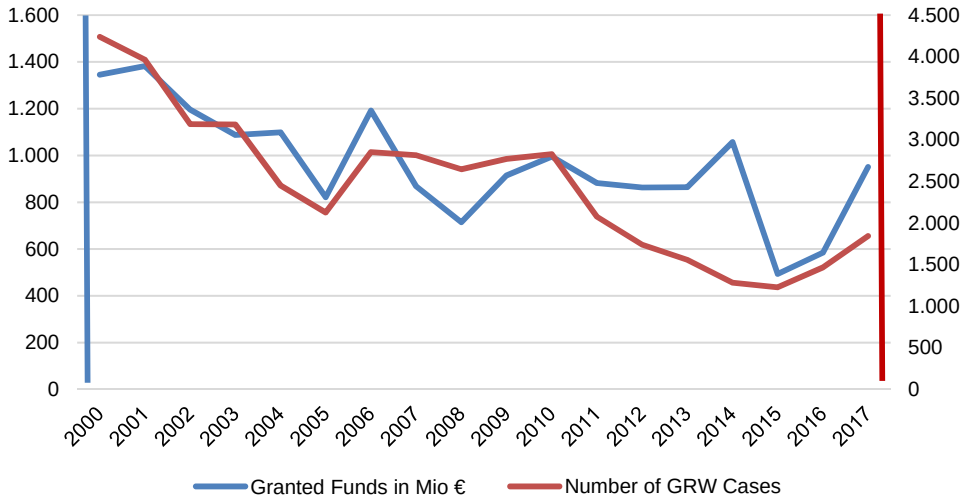
Appendix

Table 1.8: Detailed Summary Statistics of the Estimation Sample

	N	Mean	Median	SD
Cloud computing	8540	0.21	0	0.41
Number of employees	8540	423.7	75	4434.5
Sales	8540	135478.4	10044.2	2288269.5
% of employees with internet connection	8540	55.1	50	33.3
% of employees with mobile internet connection	8540	18.1	10	22.6
Broadband speed				
Below 2Mbit/s	8540	0.042	0	0.20
Between 2 Mbit/s and 10 Mbit/s	8540	0.26	0	0.44
Between 10 Mbit/s and 30 Mbit/s	8540	0.26	0	0.44
Between 30 Mbit/s and 100 Mbit/s	8540	0.21	0	0.41
More than 100 Mbit/s	8540	0.18	0	0.39
Age	8540	28.3	22	22.6
Eligibility for GRW funding	8540	0.34	0	0.47
Employment of own IT-staff	8540	0.50	1	0.50
Population density	8540	1.13	0.74	1.07
Industries				
Manufacturing	8540	0.54	1	0.50
Utilities	8540	0.024	0	0.15
Wholesale and retail trade, repair of motor vehicles ...	8540	0.100	0	0.30
Transportation and storage	8540	0.024	0	0.15
Accommodation and food service activities	8540	0.056	0	0.23
Information and communication	8540	0.11	0	0.31
Real estate activities	8540	0.031	0	0.17
Professional, scientific and technical activities	8540	0.038	0	0.19
Administrative and support service activities	8540	0.075	0	0.26
Repair of computers and communication equipment	8540	0.0044	0	0.067
Details cloud computing				
Cloud e-mail	8517	0.081	0	0.27
Cloud finance software	8515	0.050	0	0.22
Cloud data bases	8509	0.061	0	0.24
Cloud storage	8514	0.11	0	0.32
Cloud computing power	8494	0.049	0	0.22
Cloud office software	8515	0.053	0	0.22
Cloud CRM software	8508	0.052	0	0.22

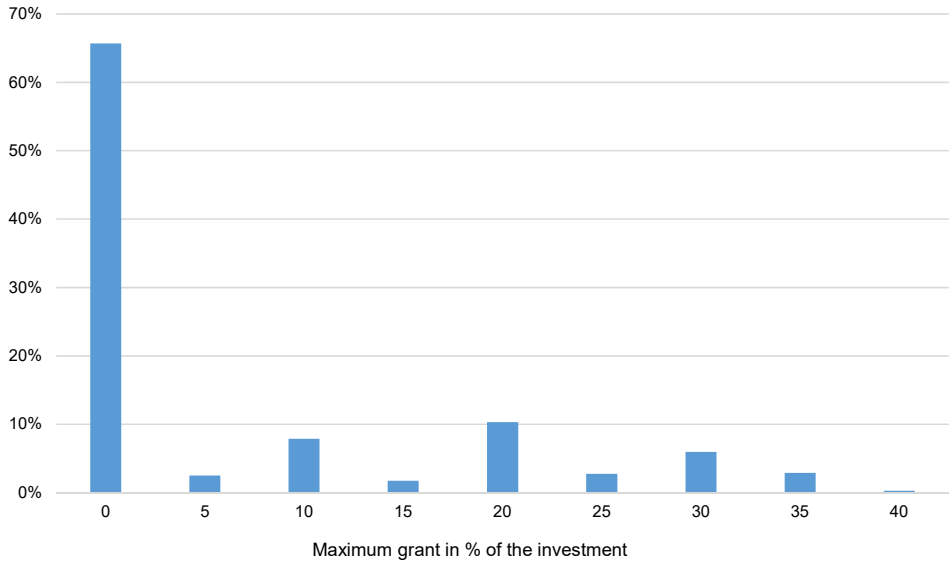
Notes: See Section 1.3 for a description of the data sources.

Figure 1.3: Development of GRW Grants



Notes: Left scale: Granted funds in Mio€(blue). Right scale: Number of GRW cases (red).
Illustrated is the number of GRW cases as well as the total sum of GRW grants awarded by year. There has been a steady decline in total grants and the number of GRW cases from 2000 to 2015. However, both figures recovered from 2015 on.
Source: Authors' calculations based on BAFA 2019.

Figure 1.4: Percentage of Firms by Eligibility of Maximum Grant



Note: This figure shows the distribution of the maximum funding rates of the GRW as determined by location and firm size over the firms in the sample. For the empirical analysis, the maximum funding of 200.000€ for large enterprises (compare to Table 10) is coded as an incentive rate of 5%. The percentages do deviate from Table 1.1 as there was a temporary increase by 5 percentage points (see Deutscher Bundestag 2014, page 16).
Source: Own illustration by authors based on ICT survey and BAFA 2019.

Chapter 2

Combating Online Hate Speech: Evidence from NetzDG

Published as

Andres, Raphaela and Olga Slivko (2021), Combating Online Hate Speech: The Impact of Legislation on Twitter, ZEW Discussion Paper No. 21-103, Mannheim.

Student Paper Award by the International Telecommunications Society 2021

2.1 Introduction

Social media have become primary information channels for many individuals (Pentina and Tarafdar 2014). In 2019, one in three people in the world used social media platforms.[19] The far-reaching spread of social media provides new opportunities for sales (Chevalier and Mayzlin 2006), inspire product innovation (Bertschek and Kesler 2022), and provide valuable input for demand forecasting (Cui et al. 2018). However, it also opens room for the dissemination of extremist thoughts, aggressive or harassing content as well as misinformation and extremist views (Aral 2020).

Negative sentiments directed at specific groups proliferate across numerous segments of social media. Since Europe's refugee crisis in 2015, the online propagation of "hate speech" has been a persistent theme in public discourse. In the US, this discourse was further exacerbated by the detrimental impact of Donald Trump's tweets (Müller and Schwarz 2023). Moreover, with the onset of the COVID-19 lockdowns, researchers and media outlets documented the emergence of concentrated hateful communities, for example, in the US and the Philippines on Twitter and Reddit (Uyheng and Carley 2021). These developments are dangerous, as the use of mass media and social media for inciting hatred can cause *stochastic terrorism*, i.e. can provoke attacks by random extremists.[20] While such trends are alarming for the society, they also hinder the performance and growth of social media platforms. The prevalence of online hate on social media platforms discourages advertisers, as they seek to distance their brand's names from hate speech and any form of abusive or dangerous content.[21]

In response to the dissemination of online hate, many platforms have introduced community standards and house rules. Additionally, they have set up infrastructures involving human moderators who collaborate with artificial intelligence-based tools. While the approaches differ substantially across platforms, regulatory bodies are paying increasing attention to this issue. The first legal framework to address online hate was implemented in Germany, which introduced the Network Enforcement Act (NetzDG) in January 2018. This law obliges large social media platforms in Germany to establish procedures for users to report hateful content and requires

[19] Our World in Data: `https://ourworldindata.org/rise-of-social-media`

[20] Wired Article: `https://www.wired.com/story/jargon-watch-rising-danger-stochastic-terrorism/`; Original Quote: `https://www.dailykos.com/stories/2011/1/10/934890/-`; Recent example: `https://www.bbc.com/news/world-europe-58635103`

[21] Vox Article: `https://www.vox.com/technology/2023/3/23/23651151/twitter-advertisers-elon-musk-brands-revenue-fleeing`

that social media platforms promptly remove respective postings. The NetzDG spurred global discussions on regulating harmful content and was used as a blueprint for similar laws in other countries (Tworek and Leerssen 2019), including the forthcoming European Digital Services Act.

We exploit the implementation of NetzDG in Germany as a quasi-experiment to measure its causal impact on the prevalence of online hate speech. To measure hate speech, we employ pre-trained algorithms provided by Jigsaw and Google's Counter Abuse Technology team which have demonstrated quite accurate performance in previous studies (Mondal, Silva, and Benevenuto 2017, ElSherief et al. 2018, Han and Tsvetkov 2020). The Application Programming Interface (API) "Perspective" allows us to measure different characteristics of short texts, including toxicity, identity attacks and profanity. Given that the application of NetzDG is restricted to the content accessible to users within German territory, we adopt a difference-in-differences framework and compare the evolution of the content produced by similar subgroups of users in the German and Austrian Twittersphere. As the two neighbouring countries share many cultural facets and the same primary language (German), this method allows us to isolate the causal effect of the regulation.

We show that NetzDG reduces the intensity of hate speech in German sensitive tweets (tackling migration and religion topics) by about 2 percentage points, which corresponds to a reduction of 6%-11% in the mean hate intensity. The volume of original hateful tweets is reduced by 11%, implying one less attacking tweet by each user in three months. While Jiménez Durán, Müller, and Schwarz 2022 also document the effectiveness of NetzDG in reducing online hate speech using a different source of variation, our paper additionally delves into the mechanisms of how NetzDG impacts the contents on Twitter. Importantly, we find no change in user engagement with these tweets as well as no change in other observable tweet characteristics, such as use of images or references to media outlets. These effects are remarkable and contribute to the current discussion on whether legal regulation of online content can effectively complement the platform's own guidelines such as the "Twitter hateful conduct policy". In a survey on NetzDG sent out to the platforms, Twitter claims that the law has not increased the deletions on the platform, as most of the illegal acts defined in NetzDG were already captured by its house rules (Liesching et al. 2021). However, this is not supported by our findings, which show that while the platform guidelines apply to both German and Austrian users, hateful content in tweets posted

in Germany significantly decreased after implementing the law compared to Austria. Notably, we find that the decrease in hate is driven by tweets tackling sensitive topics of migration and religion. For other tweets in our 2 mln. sample, there was no evidence of change in hate intensity or user engagement.

The findings from our study close important knowledge gaps for platform managers and policy makers to understand potential impacts of hate speech regulation. First, content moderation does not appear to hamper user engagement or the network effects that attract advertisers. On the contrary, since social media's advertisers care about their brand's reputation, a decrease in online hate could enhance the appeal of social media platforms for advertisers Madio and Quinn 2023. Hence, if platforms can demonstrate a decrease in hate speech, they could boost their attractiveness for advertising and potentially increase their revenues. Furthermore, we demonstrate that establishing content moderation infrastructures following the introduction of the Digital Services Act (DSA) can potentially decrease the prevalence of hate on sensitive topics.

2.2 Literature

2.2.1 The Impact of Social Media

Previous literature has established a strong connection between social media and business operations management. Gu and Ye 2014 emphasize the pivotal role of social media in fostering relationships with customers. By connecting with customers on social media, businesses can enhance satisfaction among the most valued customer segments (Magids, Zorfas, and Leemon 2015). Interactions between business and customers on social media are a valuable source of product innovations (Bertschek and Kesler 2022). Additionally, content produced by potential customers on social media has a more profound impact on product purchases compared to content developed by marketologists (Goh, Heng, and Lin 2013). Related to that, the data on interactions of businesses with customers on social media improve the accuracy of daily sales forecasts (Cui et al. 2018, Lau, W. Zhang, and Xu 2018).

Social media have also been shown to mediate political (dis)engagement and individual well-being. Allcott and Gentzkow 2017, Zhuravskaya, Petrova, and Enikolopov 2020, Enikolopov, Makarin, and Petrova 2020 and Vosoughi, Roy, and Aral 2018 highlight the role of social media in the spread of political misinformation, polarization, and as a means of political coordination.

A survey among German internet users in 2023 shows that almost 60% of the users do not share their political opinion and are less likely to participate in political online discussion due to fear of provoking hateful comments.[22] Furthermore, engaging with social media strongly affects individual well-being. Allcott, Braghieri, et al. 2020 draw a link between a temporal social media deactivation and improved subjective well-being, as well as reduction in news consumption and political polarization. Braghieri, Levy, and Makarin 2022 analyze the effect of the staggered rollout of Facebook across US campuses on student mental health and found that the negative effect of social media adoption was stronger for students that might suffer from unfavourable social comparisons.

A recent strand of studies identifies an important link between xenophobic attitudes expressed on social media and offline hate crimes (Jiménez Durán, Müller, and Schwarz 2022, Müller and Schwarz 2021, Müller and Schwarz 2023, Bursztyn et al. 2019, Olteanu et al. 2018). Müller and Schwarz 2021 measure the short-run effect of social media on violent crimes. They show that the effect of anti-refugee sentiments posted on Facebook disappears on the days of Internet outages and disruptions to Facebook access in Germany. Bursztyn et al. 2019 measure long-term effects of social media penetration in Russia on anti-immigrant hate crimes. Additionally, Olteanu et al. 2018 show that offline violence (Islamist attacks) causes online hate speech against muslims across social media platforms. Additionally to the strong connection between online and offline hate, Beknazar-Yuzbashev et al. 2022 show that toxic UGC is contagious and users exposed to lower toxicity reduce their own toxicity in posts and comments on Facebook and Twitter.

Hence, policy makers, platform stakeholders and civil society are increasingly acknowledging the importance of moderating harmful content, while also recognizing the potential side effects (e.g. censorship and limitation to the freedom of speech). Our paper contributes to this debate by presenting empirical evidence on the effects of harmful content moderation on social media imposed by the German government regulator. Our findings suggest that the harmful effects of online hate can be mitigated if platforms are legally obliged to promptly address user-reported hateful content. Furthermore, we demonstrate that only user-generated content tackling sensitive topics is affected by the regulation. We find no effect on other topics or on the users' patterns of expression on Twitter.

[22] Link to study: `https://kompetenznetzwerk-hass-im-netz.de/lauter-hass-leiser-rueckzug/`

2.2.2 Content Moderation and Regulation on Platforms

Theoretical studies on content moderation suggest that the incentives of the social media platforms to provide the optimal level of content regulation may be insufficient (Buiten, Streel, and Peitz 2020, Liu, Yildirim, and Z. J. Zhang 2022). In fact, it might be optimal for platforms to keep extreme content on the platform to extend their user base and advertising-driven profits (Liu, Yildirim, and Z. J. Zhang 2022). Additionally, the design of the regulation also matters. Feher 2023 shows that a uniform regulation that treats all platform users in the same way could encourage platforms to punish only users with low overall impact on the platform in order to avoid sanctions while keeping the ones with high impact. He highlights the importance to consider the harm that concrete platform users may cause due to their audience sizes in the regulation design.

In recent years, social media platforms increasingly undertake efforts to set boundaries on misinformation and the prevalence of harmful content. For misinformation, platforms experimented with the implementation of nudges as well as peer content moderation mechanisms. Wang, Pang, and Pavlou 2021 show that the enforcement of identity verification on a social media platform can decrease the propensity to post fake news exploiting data from Weibo, an equivalent of Twitter in China. Ershov and Morales 2021 analyze how Twitter users responded to the user interface change nudging users into adding a comment on the content they were going to share. After this change, content sharing was significantly reduced, and while there was no difference for low and high factualness, left-wing media experienced a very high drop in sharing compared to the right-wing media.

Several studies assessed platform governance mechanisms for content moderation. Chandrasekharan et al. 2017 study an event of banning two hateful communities on Reddit and show that after the ban some users reduced their usage of hate terms, while others left the platform. Srinivasan et al. 2019 analyze the evolution of swear words and hate terms within a subreddit and find no relationship between content removal and the use of hate terms for non-compliant users. Borwankar, Zheng, and Kannan 2022 assess the impact of the Birdwatch program on Twitter and show that peer content moderation increases cognition in writing and decreases content extremity at the cost of substantially decreased content quantity.

Our paper adds to these studies by providing evidence on the effects of regulation of harmful content on a social media platform. It particularly contributes to the emerging studies discussing and evaluating the consequences of the implementation of NetzDG.

2.2.3 The Impact of NetzDG

Most previous studies on NetzDG provide descriptive evidence adopting the legal and media perspective (Kasakowskij et al. 2020, Liesching et al. 2021). Several studies rely on the data from the NetzDG transparency reports published by social media platforms (for an overview see Griffin 2021). Kasakowskij et al. 2020 conclude that the vast majority of user reports on Twitter did not lead to deletion or blocking, because most of the content reported by users was, apparently, not unlawful. Also, Liesching et al. 2021 observe only a "[..] marginal importance of the Network Enforcement Act in application practice" (translated from Liesching et al. 2021, p. 368). However, the data from transparency reports are not very informative about the causal effect of NetzDG because they only include UGC reported by platform users and do not capture the totality of online hate on the platforms (Griffin 2021). Furthermore, the take down numbers within the transparency reports are likely biased since platforms have an incentive to delete under house rules instead of NetzDG to avoid legal consequences (Echikson and Knodt 2018). This incentive is mirrored by a survey sent to the platforms, in which the platforms claim that the increased deletion practice documented in the NetzDG transparency reports is due to the platform's house rules and not due to NetzDG (Liesching et al. 2021). Contrary to these studies, our paper uses data directly drawn from one of the largest social media platforms and analyzes the effect of the law on UGC in a quasi-experimental setting. Comparing UGC in the treated and the control groups, we show that there is an additional reduction in online hate due to regulation in the presence of the platform's own governance mechanisms.

For the more causal assessment of the impact of NetzDG, recent studies rely on quasi-experimental and experimental approaches. In a large field experiment, Jiménez Durán 2022 addressed the impact of NetzDG on the likelihood of removing the reported content by Twitter. The author finds that reported tweets are more likely to be deleted (3.5%) while non-reported hateful tweets are only 2.1% likely to be deleted. While Jiménez Durán 2022 suggests no evidence of self-censorship for the users whose tweets were deleted, the decrease in toxicity in our setting is driven by users decreasing hate intensity in their tweets on sensitive topics due to

self-censorship. Complementary to our study, Jiménez Durán, Müller, and Schwarz 2022 find that the NetzDG is associated with a reduced toxicity in the German right-wing Twitter segment. Their finding reaffirms our results of the law's effectiveness by exploiting a different source of variation: While Jiménez Durán, Müller, and Schwarz 2022 use within-country variation by comparing followers of the right-wing party to less affected followers of other parties, our paper compares Twitter users located in Germany to Twitter users located in Austria who unaffected by the regulation (cross-country variation). Furtermore, Jiménez Durán, Müller, and Schwarz 2022 combine their finding with localized Facebook user data and demonstrate that the reduced online toxicity also leads to decreased offline violence, emphasizing the offline impact of NetzDG. Our paper contributes to the analysis of NetzDG's effectiveness by delving into the mechanisms of *how* NetzDG affects the content generation and user engagement on social media.

2.3 Overview of NetzDG

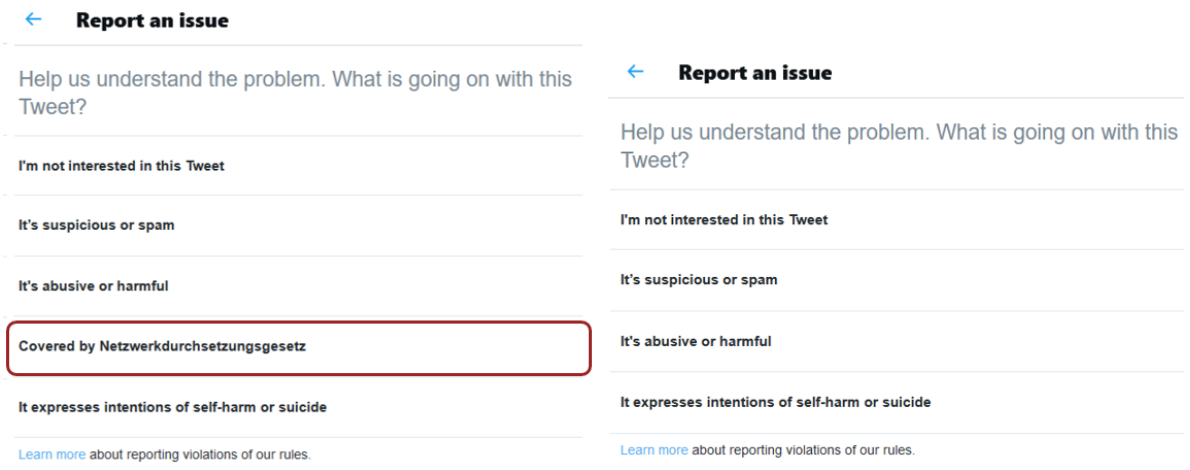
NetzDG[23] was passed by the German Bundestag in October 2017 and came into effect in January 2018. The law aims at reducing the prevalence of online hate speech by increasing the legal pressure on platforms to act against hateful content generated by users. Specifically, it obliges social media platforms with more than two million registered users in Germany[24] to implement mechanisms that provide each user with a transparent and permanently available procedure to report illegal content on the respective platform. After receiving a complaint, the platform is required to review the complaint immediately and act within a reasonable time frame. If the user complaint targets unquestionably illegal content, it must be removed within 24 hours. In more nuanced cases, platforms have seven days to decide whether measures must be taken against the respective content or the user account which submitted it. In practice, Twitter decided to add the option "Covered by Netzwerkdurchsetzungsgesetz" if the user accessed Twitter via a German IP address (see Figure 2.1).[25]

Next, the reporting users need to choose the paragraph of the criminal code violated by the post. Finally, they must sign an acknowledgement that the wrongful reporting of a tweet itself is

[23] Netzwerkdurchsetzungsgesetz, for English version of the law see: `https://germanlawarchive.iuscomp.org/?p=1245`

[24] As of December 2020, this applies to: Facebook, Youtube, Instagram, Twitter, Reddit, TikTok, Change.org, Jodel (BMJV 2020)

[25] If users located in Germany click on the broader option "It's abusive or harmful", he or she can indicate "Covered by Netzwerkdursetzungsgesetz", while other options are to report the usage of private information and incitement of suicide or self-harm.



(a) German IP address (b) Austrian IP address

Figure 2.1: Menu Options for Reporting Tweets with a German/Austrian IP Address

a violation of the Twitter house rules. After this procedure, the platforms can choose which steps to undertake to address the complaints: they can remove the content in case it is clearly hateful, send a warning to the user account that posted it, or temporarily or permanently block this user account.

Importantly, the law does not require platforms to proactively search and delete hate speech, but only to become active after receiving a concrete complaint that indicated the "Covered by Netzwerkdurchsetzungsgesetz"-option. Further NetzDG requirements include the appointment of a domestic authorized contact person for each platform and the semi-annual publication of the compliance report. This report should include information on how to report illegal content and the total numbers of content removal requests by groups of users (private users or organizations), reaction time, and the reason for reporting.

According to §4 of NetzDG, non-compliance can be penalized with a fine of up to five million €. However, due to the risk of content overblocking, the examples for punishable offences only include technicalities about the report and a systematic incorrect execution or monitoring of the complaint management system. To prevent platforms from pursuing the “better to be safe than sorry” strategy and delete any content that might seem questionable at first sight, social media

platforms are deemed non-compliant only if they fail systematically to meet the requirements of the regulation.[26]

2.4 Data and Empirical Strategy

2.4.1 Data

We measure the effect of NetzDG on tweets posted by followers of the right-wing populist party Alternative für Deutschland (AfD), which is represented in the German Bundestag. This segment of the Twittersphere is particularly relevant for our research question because hate speech has a higher prevalence among these users. As studies suggest, individuals with populist views are more likely to use strongly negative rhetoric towards different social groups, e.g. migrants (Halikiopoulou 2018), hence, to fuel hate speech. The xenophobic content generated by right-wing users in Germany directly connects to incidences of hate crime (Müller and Schwarz 2021). Therefore, the impact of the law on this part of the Twitter community is of particularly high interest.

For our analysis, we observe all national and regional profiles of the AfD party on Twitter which we could find at the time of data extraction (May 2020) and ended up with 201 national and regional AfD profiles. However, since official party profiles need to consider the phrasing of their tweets more carefully than an individual user and we want to estimate the effectiveness of NetzDG on a broader public, we downloaded all followers of those party profiles and focus on the contents posted by those users. Importantly, only users with a German IP address are affected by NetzDG. Therefore, we reduced our sample to users who indicated living in Germany according to their profile information. In our full sample, 63% of the users provided some information about their location in their profile such that they could cleanly be assigned to the treatment group (being located in Germany) versus the control group (being located outside Germany). The resulting sample is a bit more active and better connected on Twitter compared to those users not indicating any location. Hence, our sample is composed of the majority of the users and measuring the performance of NetzDG is even more important on this subgroup. For these

[26] Under NetzDG, Facebook was fined five million € for an erroneous compliance report (`https://www.heise.de/news/NetzDG-Verstoesse-Facebook-hat-fuenf-Millionen-Euro-an-Strafen-gezahlt-6181705.html`) and Telegram for not setting up a suitable reporting procedure and a domestic authorized complaints recipient (https://www.tagesschau.de/inland/innenpolitik/strafe-telegram-fuenf-millionen-101.html). These are the only fines under NetzDG as of March 2024.

followers, we observe all their original tweets between July, 2016 to June, 2019 (1.5 years before and 1.5 years after the introduction of NetzDG) which were still present on the platform at the moment of data collection in May 2020. Throughout the analysis, we only consider original tweets, as opposed to retweets, because we focus on the language that users in our sample choose. Finally, out of 2.3 million retrieved tweets, we selected tweets tackling the topics of migration and religion[27] in messages and hashtags as German-language tweets related to anti-immigrant and anti-muslim topics are the most likely to contain hatred according to the "Political Speech Project"[28].

As a control group, we chose a similar German-language segment of Twitter which was not affected by NetzDG: we could manually identify about 30 profiles of the right-wing populist party in Austria, Freiheitliche Partei Österreichs (FPÖ). Out of all followers of this party, we selected all users who are *not* located in Germany and are therefore not treated by NetzDG.[29] Importantly, the language used by German and Austrian users is similar, as German is the mother tongue in both countries. Although the spoken Austrian German sounds like a dialect to German users, written German is the same in both countries. Furthermore, even if there were slight differences between the language of German and Austrian users, these level differences would be differenced out in our estimation approach. Hence, our control group allows us to analyze the development of hate speech in two comparable segments of Twitter before and after the implementation of NetzDG.

Our resulting data set comprises more than 160,000 tweets for German and Austrian right-wing followers about sensitive topics, like migration and religion, as our sample for the baseline analysis. Importantly, our sample composition implies that we can *not* measure the effect of the law for the entire Twittersphere, but rather for an important target group of the law. Due to the negative real-world consequences of hate speech posted by right-wing populists (Caiani and Parenti 2013), the effect of the law on this segment on Twitter is of particularly high interest.

[27] For the filtering, we used the following word stems: reli, migra, islam, terror, flucht, flücht, moslem, koran, ausländ, ausland

[28] `https://rania.shinyapps.io/PoliticalSpeechProject/`

[29] In a robustness check, we only considered users indicating living in Austria as a control group and all results stay the same.

2.4.2 Outcome Variables

We measure the intensity of hate in tweets using Jigsaw and Google's Perspective API, which employs pre-trained machine learning models to score the probability that short texts are hateful. In the natural language processing literature, Perspective API is considered a benchmark prediction algorithm (Fortuna, Soler, and Wanner 2020).[30] It relies on a convolutional neural network trained on large corpora of publisher and user-generated content from multiple domains (such as Wikipedia, the New York Times, The Economist, The Guardian, including user comments on their forums).

Since the NetzDG itself does not include any measurable definition of hate speech, we use several dimensions of hate speech available in Perspective API for the German language, namely, severely toxic, toxic, threatening, an identity attack, profane, and insulting language. Exploring these different dimensions allows us to learn more about potential channels through which the law might tackle the issue of hate speech. As our outcomes for both treatment and control group are tweets written in the German language, we evaluate hate intensity in both groups using the *same* algorithm. Therefore, potential prediction biases are distributed randomly across tweets in our sample and do *not* affect our results due to our identification strategy.

Perspective API algorithm evaluates the probability scores of each tweet to contain hate for each of the six dimensions in the range $[0, 1]$, so that the probabilities can be interpreted as intensities of hate in tweets. In our analysis, we multiply these scores by 100 to improve the interpretation of the estimation coefficients.

2.4.3 Summary Statistics

Following the extraction procedure described in Section 2.4.1, we obtained 735 right-wing sympathizers located in Austria and 602 users in Germany. Several users (187) indicated that they did not live in Germany or Austria in their profile information. Since we are not interested in the user's residency per se but only if they live in German territory and are therefore exposed

[30] We additionally tested the performance of the algorithm on a random subset of 100 tweets by comparing the estimated scores of all six hate dimensions to scores that were provided by four human classifiers. While the distributions of the scores along the unit interval seem to be similarly captured by the humans and the algorithm (see Figure 2.5 in the Appendix), the tweets that were defined as very hateful differ between all classifications. Not only is the overlap between the majority of humans and the algorithm rather low, also the agreement among the humans is low. We therefore conclude that hate speech recognition remains a difficult task and rely on the Perspective API as the workhorse model in the natural language processing literature.

to the regulation, we assign those users to the control group together with the users from Austria. We kept an indicator for those profiles to account for potential differences in tweets between those living in Austria and those living somewhere else and tested for the robustness of our results by dropping these users. Table 2.1 presents measures describing the profiles of users in our sample. Most of the user characteristics in Table 2.1 are quite dispersed. For example, the number of followers ranges from 0 to almost 550,000. The oldest profile in our sample was created in 2007, whereas other users created their accounts after the introduction of NetzDG. Some users (18%) only tweeted once. This might be due to low account age, inactivity during our sample period, or little interest in migration or religion, since we only include tweets about these sensitive topics in our main sample. Among our randomly chosen accounts, 22 (1.6%) are the user accounts of politicians (i.e., members of the German or Austrian parliament), and 28 accounts belong to a well-known personality ("verified").

Table 2.1: Summary Statistics of User Characteristics

	N	Mean	Median	SD	Min	Max
No. of Followers	1334	2008.91	236	16337.57	0	534819
No. of Friends	1334	1601.41	487	16668.30	1	590754
Year of account creation	1335	2014.07	2014	3.00	2007	2019
Verified account	1335	0.02	0	0.14	0	1
Live in GER	1337	0.45	0	0.50	0	1
Live outside GER/AUT	1337	0.14	0	0.34	0	1
Only 1 tweet in sample	1337	0.18	0	0.39	0	1
No. of tweets in sample	1337	120.03	9	524.00	1	12279
No. sens. tweets user/month	1337	10.09	2	33.80	1	848
Politician	1337	0.02	0	0.13	0	1

Notes. The table shows summary statistics on the user level. All statistics combined show that the users in our sample are diverse with regard to Twitter activity and connectedness.

Table 2.2 presents summary statistics on the tweet level. Besides the tweet text, we extracted additional meta information such as the number of retweets, likes, and replies. The popularity of tweets can differ greatly. Most tweets are not retweeted or liked, whereas others have more than 1,700 likes. The median tweet length in our sample comprises 15 words, while the number of possible characters of a tweet doubled from 140 to 280 characters within our sample period. As Twitter imposed this rule for both countries simultaneously and we include month fixed effects in every estimation, the increase of allowed characters does not threaten the validity of our

identification strategy. Further information we collected on the tweets are indicators if the tweet includes a video, photo, URL, or a link to a media outlet. We also observe the time when the tweet was posted and added a country-specific daily indicator if a terrorist attack or an election (European, national or regional elections) took place in Germany or Austria. Within our sample period, national elections in Germany as well as in Austria took place in the fall of 2017.

Table 2.2: Summary Statistics of Tweet Characteristics

	N	Mean	Median	SD	Min	Max
Severe Toxicity	160474	29.98	29	24.13	0	100
Toxicity	160474	42.81	45	22.47	0	100
Threat	160474	34.73	21	24.90	0	100
Identity Attack	160474	57.42	61	27.59	0	100
Profanity	160474	20.48	11	20.47	0	100
Insult	160474	37.24	36	22.06	0	100
No. of Retweets	160474	4.00	0	19.35	0	911
No. of Likes	160474	7.03	0	36.12	0	1711
No. of Replies	160474	1.12	0	5.62	0	292
Video in tweet	160474	0.00	0	0.05	0	1
Photo	160474	0.07	0	0.26	0	1
URL	160474	0.68	1	0.46	0	1
Link to media outlet	160474	0.06	0	0.24	0	1
No. of Words	160474	18.19	15	9.60	1	57
Tweeted at night	160474	0.08	0	0.26	0	1
Terrorist attack in country	160474	0.02	0	0.12	0	1
Election in country	160474	0.01	0	0.10	0	1

Notes. The table shows summary statistics on the tweet level. The first rows are the outcome variables of the main analysis. Subsequently listed are tweet characteristics such as the number of retweets and number of words. Lastly, we included country-specific indicators for days when an election and/or terrorist Attack took place.

In Table 2.3 we compare the average of all outcome dimensions between the treated and control group. The overall intensity of hateful content is higher in our sample of German compared to Austrian users. However, the descriptive evidence suggests that in Germany, the mean values decreased after NetzDG became effective, whereas they increased in Austria. Table 2.16 (see Appendix) shows the pairwise correlations among the outcome variables, indicating high correlations between toxicity and insults and between severe toxicity and toxicity.

Table 2.3: Outcome Variables by Country and before/after

	Germany before	Germany after	Austria before	Austria after
	Mean	Mean	Mean	Mean
Severe Toxicity	33.4	28.7	28.4	28.3
Toxicity	45.4	42.0	40.3	42.8
Threat	37.6	34.8	33.0	31.9
Identity Attack	59.9	57.4	52.8	58.6
Profanity	20.8	20.2	19.1	21.8
Insult	38.1	37.0	34.6	39.2
Observations	47855	49281	33016	30322

Notes. The table shows the average of all hate dimension scores by country and before/after NetzDG became effective.

2.4.4 Empirical Model

Since the application of NetzDG is restricted to the content on social networks on the German territory, we apply a difference-in-differences (DID) framework comparing the evolution of the language used by comparable subgroups of users of the German and Austrian Twittersphere. We estimate the DID by ordinary least squares (OLS) and include fixed effects (FE) for users, calendar months, and account age at the time the tweet was posted:

$$Hate\ Intensity_{ijt} = \beta_0 + \beta_1\ AfterT_t\ Treated_{ij} + X_{it}^{'}\ \beta_2 + \mu_j + \nu_t + k_{t'} + \varepsilon_{ijt}$$

We estimate separate regression models for each of the hate speech outcomes, such that the left-hand side of the equation $Hate\ Intensity_{ijt}$ corresponds to the respective hate intensity of a tweet i issued by user j on day t concerning severe toxicity, identity attacks, etc. provided by Perspective API. $X_{it}^{'}$ is a vector of time variant control variables indicating the day of the week the tweet was posted and if the tweet was posted at night. We also added country-specific daily indicators for terrorist attacks, and national or regional elections, as these events could affect the usage of hate speech in a country which would not be captured by country fixed effects. In both of the countries, national elections took place in the fall of 2017. The coefficient of $AfterT_t\ Treated_{ij}$, β_1, is the coefficient of interest, which measures the change in the hate intensity in a tweet in Germany after NetzDG. μ_j represent user FE to control for user-specific tweeting style and ν_t account for calendar month FE to capture general time trends. We additionally include account age FE $k_{t'}$, as the literature suggests that cohorts of social network

users may differ in their writing style (Ershov and Mitchell 2020). ε_{ijt} indicates the stochastic error term.

2.5 Results

2.5.1 The Effect of NetzDG on Hate Intensity

Table 2.4 reports the results of our baseline specifications.[31] The results suggest that the intensity of hate speech significantly decreases after the introduction of NetzDG in Germany. As our dependent variable is measured on a scale between 0 and 100, the coefficients of interest are interpreted as percentage point (pp) changes in the dependent variables. Hence, Table 2.4 shows that NetzDG significantly reduces the intensity of severe toxicity, toxicity, and insulting remarks by 2 pp and profanity by 1 pp. Noteworthy, the introduction of NetzDG has the highest effect on tweets related to identity attacks: the probability significantly decreased by 3 pp.

The comparison of the effect sizes to the means of the outcome variables hints at modest effect sizes. The average intensity of an identity attack in all tweets in our sample is 57. At the mean, this would decline by 3 pp and result in an average intensity of 54. In percentage terms, these numbers indicate a reduction in hate intensity of 5% of the mean value or 9% of the standard deviation. Similarly, for severe toxicity the decline is 2 pp, which implies a reduction in hate intensity by 6% of the mean or 8% of the standard deviation. The changes in hate intensity are highly significant for all of the dependent variables - except for threat intensity, which is insignificant throughout our analysis. This can be explained by the fact that threats have already been actionable and illegal before NetzDG. Moreover, the evaluation of Perspective API in Fortuna, Soler, and Wanner 2020 and also in our manual check (see Figure 2.5 in the Appendix) suggest that the classifier performs worse at identifying threats.

These results remain strong and robust to sample composition tests. Using a balanced sample consisting of accounts that tweeted before and after NetzDG and excluding users living outside of Germany and Austria does not alter the results (see Table 2.11 and Table 2.19 in the Appendix). Furthermore, omitting the transition period (i.e., six months before the introduction of NetzDG which elapsed between the moment when the law was approved by the Bundestag and actually

[31] Table 2.17 in Appendix presents the full list of control variables with the respective coefficients. This result was presented at the ICIS conference on December 13th, 2021 (Andres and Slivko 2021b). All other results are novel contributions.

Table 2.4: Baseline Analysis: The Effect of NetzDG on the Intensity of Hate in Tweets

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID Attack	(5) Profanity	(6) Insult
Treated after T.	-1.89***	-2.15***	-0.76	-2.63***	-1.33***	-2.18***
	(0.68)	(0.54)	(0.71)	(0.81)	(0.50)	(0.55)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.16	0.18	0.07	0.20	0.12	0.18
Observations	160165	160165	160165	160165	160165	160165
Mean of Outcome	29.97	42.81	34.73	57.42	20.48	37.24
SD of Outcome	24.13	22.46	24.90	27.59	20.47	22.06

Notes. The table shows the main coefficients of the difference-in-differences estimations comparing the hate intensity in tweets by users affected and unaffected by the law (NetzDG). The columns contain the outcome measures discussed in the data section: Continuous scores ranging from 0 to 100 with regard to severe toxicity, toxicity etc. as calculated by Perspective API. The coefficient *Treated after T.* shows the change in hate intensity in terms of percentage points for users located in Germany after NetzDG became effective. Besides the treatment effect, all estimations control for country-specific events of regional/national elections and terrorist attacks, the day of the week the tweet was sent and an indicator if the tweet was sent at night. All estimations include a constant and year-month fixed effects, user fixed effects, and fixed effects for the account age in months when the respective tweet was posted. Standard errors are clustered at the user level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$.

came into force, and which also includes the national elections in both countries) also does not change the results (Table 2.20). Moreover, our preferred specification only controls for the weekday and indicators for night-time, terrorist attacks, and elections, since we consider the tweet characteristics shown in the summary statistics (Table 2.2) rather as potential outcomes of the treatment effect. However, including these tweet characteristics as controls in a robustness check does also not affect our results and further confirms the robustness of our baseline specification. Lastly, our results are robust to the placebo treatment. If we set the moment of the law implementation to January 2017, the year before the actual implementation, the treatment effect vanishes as expected (see Table 2.21).

Our results are based on the assumption that the measures of hate speech followed comparable trends in the treated and control group before NetzDG was introduced in Germany. We test the parallel trend assumption by decomposing the treatment effect by quarters before and after the regulation was introduced. Figure 2.6 (in Appendix, replicated from Andres and Slivko 2021b) presents the results for our six dependent variables of interest, corresponding to Col. (1) - (6) in Table 2.4. The standard errors of coefficients plotted in the figures correspond to the 90%

significance level. The graph shows that the treatment and control groups do not systematically differ *ex ante*, but they do differ in the quarters subsequent to the treatment (except for panel (c) (threat), for which we do not find any effect of the regulation).

2.5.2 Identification

Our identification strategy allows us to measure the causal effect of the regulation on the intensity of hate in tweets. As the parallel trends assumption suggests, there were no significant differences in the trends of hate before the regulation between the German and the Austrian Twitter segments. Even if there would be differences in the levels of hate, due to, for example, API measuring Austrian language specificities differently from the German language, differencing out the levels allows us to focus on the changes.

We could further be concerned that there is contamination between our treated and control Twitter segments. However, our design compares followers of right-wing parties located in Germany with those who are located outside Germany. Moreover, we can exclude followers who are following both German and Austrian parties and our results are unchanged. Hence, relatively isolated segments without interaction between each other are driving our results. This releases our worry about the potential contamination between the users in the treatment and control groups.

We further address a concern about the potential heterogeneity across Twitter users in our treated and control groups using coarsened exact matching (CEM). Compared to the widely used propensity score matching, CEM does not require assumptions on the model connecting covariates and potential outcomes and helps to control the potential imbalances in the covariates (King and Nielsen 2019). CEM coarsens a set of the observed covariates, and then matches the coarsened data. For matching the Twitter users located in Germany with those located outside Germany, we use the set of covariates that describe the average patterns in the user activity in the period between July and December of 2016, a year before the discussion of the regulation went public (see Table 2.5).

Based on these covariates, 462 followers from our sample were matched with each other. For these matched followers from the treated and control groups, we again compare our hate intensity measures in tweets before and after the implementation of NetzDG. Our matched

Table 2.5: Covariates for Coarsened Exact Matching

Variable	Description
Toxicity	Average level of toxicity across all the tweets that each user posted in the period between July and December of 2016.
Insult	Average level of insult across all the tweets that each user posted in the period between July and December of 2016.
Tweeting frequency	Average monthly number of tweets in the period between July and December of 2016.
Word count	Average word count across all tweets that each user posted in the period between July and December of 2016.
Night	Share of tweets posted in the nigh time between 22pm and 7am.
Video	Share of tweets containing videos.
Retweets count	Average number of retweets per tweet for each user in the period between July and December of 2016.
Likes count	Average number of likes per tweet for each user in the period between July and December of 2016.
Verified	Indicator whether the user's Twitter account is verified. It takes value 1 if the account is verified, and 0 otherwise.
Politician	Indicator whether the owner of the Twitter account is a politician. It takes value 1 if the user is a politician, and 0 otherwise.

sample of tweets contains more than 50,000 observations. The results in Table 2.6 again suggest that due to NetzDG, the hate intensity in German tweets decreased by 2-3 pp. Similarly to our baseline specification, severe toxicity decreases by about 2 pp and the intensity of identity attacks decreases by 3 pp. Again, the intensity of threat in tweets is insignificant. Hence, our baseline findings show robustness to any potential differences in the composition of users in the treated and control group.

2.5.3 The Effect of NetzDG on the Volume of Hateful Tweets

In addition to the hate intensity in tweets, we address the effect of NetzDG on the volume of original hateful content posted by the followers of the right-wing party located in Germany. We set up a panel at the user-month level and aggregate the number of tweets containing hate speech

Table 2.6: Baseline Analysis on the Coarsened Exact Matched Sample: The Effect of NetzDG on the Intensity of Hate in Tweets

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID Attack	(5) Profanity	(6) Insult
Treated after T.	-1.96**	-2.57***	-1.59*	-2.99**	-1.84***	-2.49***
	(0.77)	(0.84)	(0.94)	(1.24)	(0.61)	(0.82)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.09	0.11	0.06	0.13	0.08	0.12
Observations	47768	47768	47768	47768	47768	47768
Mean of Outcome	27.76	42.20	32.62	57.63	20.12	37.44
SD of Outcome	23.16	22.30	23.71	27.59	19.99	21.94

Notes. The table shows the coefficients of interest in the difference-in-differences estimations comparing the hate intensity in tweets by CEM-matched users affected and unaffected by the law (NetzDG). The columns contain the outcome measures discussed in the data section: Continuous scores ranging from 0 to 100 with regard to severe toxicity, toxicity etc. as calculated by Perspective API. The coefficient *Treated after T.* shows the change in hate intensity in terms of percentage points for users located in Germany after NetzDG became effective. Besides the treatment effect, all estimations control for country-specific events of regional/national elections and terrorist attacks, the day of the week the tweet was sent and an indicator if the tweet was sent at night. All estimations include a constant and year-month fixed effects, user fixed effects, and fixed effects for the account age in months when the respective tweet was posted. Standard errors are clustered at the user level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$.

according to Perspective API. Since the outcome variables are measured in intensities of the six hate dimensions, we constructed an indicator for each tweet and defined a tweet as belonging to the category, for example, "severely toxic" if the probability of being severely toxic is above 80 - i.e., it is very likely that the tweet is severely toxic. This threshold has been recommended by Perspective API and supported by computer science research (Mondal, Silva, and Benevenuto 2017, ElSherief et al. 2018, Han and Tsvetkov 2020). Our resulting panel is very unbalanced as very few users tweet frequently about migration and/or religion. Therefore, in the following estimations we only include users who tweeted at least twice before and after the introduction of NetzDG to properly account for user fixed effects.

Our fixed effects estimations in Table 2.7 yield a similar picture to the tweet-level estimations. The coefficients with respect to all of the measures of hate speech are negative but less precisely estimated. The effects are significant for severe toxicity, toxicity, and identity attacks, which are the most discussed measures in the literature addressing automated hate speech detection (ElSherief et al. 2018, Han and Tsvetkov 2020) and are the ones for which the Perspective API demonstrates better performance (Fortuna, Soler, and Wanner 2020). Hence, the volume of

potentially unlawful tweets also declined in Germany as a consequence of NetzDG. Since we estimate the impact of the law on the logarithmic outcomes, the coefficients are interpreted as semielasticities of the change in the number of potentially unlawful tweets. According to Table 2.7, the number of identity attacks fell by 11% in Germany due to the introduction of NetzDG. Comparing this effect to the average number of identity attacks by user and month throughout the sample (3) implies that on average, there is one identity attack less per user in three months in Germany.

Table 2.7: Panel: The Volume of Hateful Tweets by User and Month in Logs

	(1)	(2)	(3)	(4)	(5)	(6)
	sev. Toxicity	Toxicity	Threat	ID Attack	Profanity	Insult
Treated after T.	-0.07**	-0.05*	-0.05	-0.11**	-0.03	-0.03
	(0.03)	(0.03)	(0.04)	(0.05)	(0.02)	(0.02)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.014	0.009	0.033	0.021	0.006	0.006
Observations	9546	9546	9546	9546	9546	9546
Groups	492	492	492	492	492	492
Mean of Outcome	0.743	0.476	1.665	3.380	0.358	0.361
SD of Outcome	2.703	1.812	5.222	8.594	1.448	1.516

Notes. The table shows the coefficients of interest of the panel difference-in-difference estimations at the user-month level. For each user and each month, the number of hateful tweets is the number of tweets with hate intensity $> 80\%$. The sample is restricted to users who posted at least twice before and after NetzDG. Besides the treatment effect, all estimations control for the country-specific share of tweets posted during night times and on days of regional/national elections and terrorist attacks. All estimations include a constant and user and year-month fixed effects. Standard errors are clustered at the user level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Interestingly, investigating the change in volume of *all* tweets that are potentially hateful and not as shown in Table 2.24 in the Appendix reveals an important differentiation: While the volume of potentially unlawful tweets is reduced by 6-12%, there is no effect of the law on the overall volume of tweets posted by the followers in our sample. This indicates that silencing discussions does not seem to be an important side effect of NetzDG: While the amount of potentially hateful postings decreases, the users do not discuss less in general. This contradicts the notion of critics of tha law that NetzDG might lead to a decrease in the freedom of speech and a silencing of discussions.

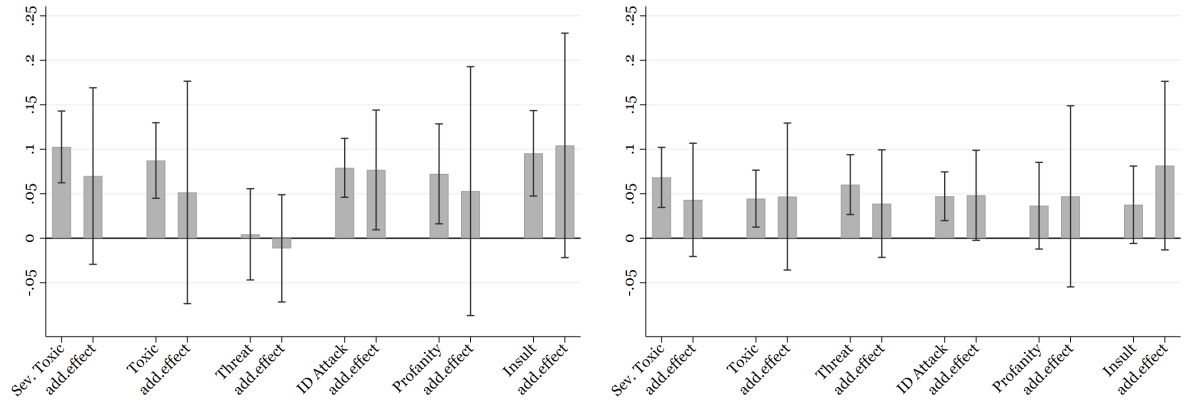
2.6 Implications of NetzDG

2.6.1 User Engagement and Effect Size

A decrease of 2-3 pp in our baseline specifications corresponds to a decrease of 5-6% in the mean hate intensity in tweets and 6-10% of a standard deviation. These numbers measure the lower bound of the effect, as our data are drawn ex post and do not include tweets by users who have been banned from the platform due to violating NetzDG. Furthermore, NetzDG concerns only the deletion of hateful content, but the overall impact from removing hateful tweets might be much larger. This is because user feeds are shaped by the algorithms which select tweets maximizing the potential attractiveness for users, based on impressions and user engagement metrics (i.e. likes, retweets and comments). Due to these algorithms, user engagement with hateful tweets increases the further exposure and subsequent user engagement with these tweets. Therefore, the law may additionally decrease hate speech on Twitter via decreased user engagement.

We examine how user engagement with tweets changes after the introduction of NetzDG. On Twitter, user engagement can be measured by the number of likes, retweets, and replies a tweet receives and greatly differs in the tweets in our sample (see Table 2.2). As in previous sections, we define a tweet as e.g., an identity attack if the score of identity attacks estimated by the Perspective API exceeds 80. To causally analyze if the user engagement with these posts changed in response to the law, we apply a difference-in-difference-in-differences (DIDID) approach. Since there was a general increase in the number of Twitter users in both countries, it is important to account for the time trends by comparing the user engagement with German and Austrian hateful tweets and other tweets before and after NetzDG.

Figure 2.2 presents the coefficients of interest for the estimation of the impact of NetzDG on the log number of “likes” and “retweets” of individual tweets, while the regression tables for all indicators of user engagement can be found in Tables 2.26 - 2.31 in the Appendix. The first bar of each color shows the coefficients of the indicator if a tweet was classified as hateful (toxic, insulting, etc.). The second bar illustrates the treatment effect for hateful tweets. Further interaction coefficients of the DIDID analyses are shown in Tables 2.26 and 2.27 in the Appendix. This analysis shows that hateful tweets receive higher user engagement, collecting significantly more likes (7%-10%) and replies (3%-5%) and are more often retweeted (4%-7%) than the non-hateful ones. This evidence is consistent with Mallipeddi et al. 2021, who show that negative



(a) Log Likes (b) Log Retweets

Figure 2.2: Coefficients Plot: Hateful Tweets Receive more User Engagement

Note: Coefficients plot of the DIDID estimation comparing the number of retweets (in logs) of hateful and non hateful tweets before and after NetzDG by treated and untreated users. The first coloured bars show the coefficient for a hateful (i.e., severely toxic) tweet while the second bars show the additional treatment effect for those hateful tweets due to NetzDG. All estimations include interaction terms "AfterT x Germany", "AfterT x Hateful", and "Germany x Hateful" and control for country-specific events such as elections and terrorist attacks, the day of the week the tweet was posted and an indicator if the tweet was posted at night. All estimations include a constant and user fixed effects, year-month fixed effects and fixed effects for the account age in month. Standard errors are clustered at the user level.

sentiments in tweets are associated with higher user engagement. However, we find no treatment effects on user engagement with hateful tweets. This suggests users do not compensate less hateful tweets by granting stronger promotion for these tweets due to NetzDG.

Since Twitter displays popular tweets on other users' feeds,[32] the significantly higher user engagement with potentially illegal tweets implies that a decrease in the number of hateful original tweets decreases the total exposure to hateful tweets overproportionally. Moreover, Beknazar-Yuzbashev et al. 2022 show in an experimental setting that toxicity is contagious. This implies that when users are exposed to less toxicity, they also reduce their own toxicity in posts and comments. Hence, the actual decrease in hate exposure due to NetzDG is higher than our baseline finding and documents the lower bound of the policy effect.

2.6.2 Content Targeting

Our baseline model (Table 2.4) shows that hate intensity decreased in Germany due to the implementation of NetzDG for tweets that address sensitive topics related to migration and religion. However, it is important to understand the broader effect of the law on the content

[32] `https://help.twitter.com/en/using-twitter/twitter-timeline`

posted by social media users in Germany. To assess this broader effect, we replicate our baseline analysis using *all* the tweets posted by the observed right-wing followers in the period from July 2016 to June 2019.

To do so, it is important to differentiate between potentially different impacts of the NetzDG on sensitive versus non-sensitive topics and conduct a triple difference analysis using all tweets posted by the right-wing followers in our sample.

$$Hate\ Intensity_{ijt} = \beta_0 + \beta_1\ AfterT_t\ x\ Treated_{ij} + \beta_2\ AfterT_t\ x\ Sensitive_i +$$
$$\beta_3\ Treated_{ij}\ x\ Sensitive_i + \beta_4\ AfterT_t\ x\ Treated_{ij}\ x\ Sensitive_i + \beta_5\ X_{it}^{'} + \mu_j + \nu_t + k_{t'} + \varepsilon_{ijt}$$

Table 2.8 provides interesting insights: While the table confirms the previous findings of NetzDG significantly reducing the hate intensity of sensitive tweets, the law does not have an impact on non-sensitive topics.[33]

Interestingly, and confirming our apporach to filter sensitive topics, the mean values of the hate scores are significantly lower in the large dataset than in the migration and religion related subsample. Furthermore, the indicator that a tweet tackles a sensitive topic is positive and strongly significant, suggesting that hate intensity in all dimensions is significantly higher in these "sensitive" tweets. These results confirm that tweets with lower average hate intensity are not affected by NetzDG suggesting that at the time of our data collection, Twitter carefully moderated content and targeted well tweets which are prone to hate speech without affecting other tweets.

2.6.3 User Tweeting Style

Our baseline findings suggest that after the implementation of NetzDG, the average hate intensity of tweets as well as the volume of tweets with high hate intensity decreased in the German Twittersphere. However, social network users might have adopted other ways to express hate, while reducing hate in the texts. For example, when the use of severely toxic language is bounded by the law, online users may express hate via hateful images or videos, or by adding links to specific media with biased articles. If social media users adjust to the regulation substituting texts

[33] Table 2.22 in the Appendix shows that when running the baseline specification again using *all* tweets posted by the right-wing followers in our sample, the impact of NetzDG vanishes. This indicates that important effects of the NetzDG would have been overlooked when not differentiating between sensitive and non sensitive topics.

Table 2.8: Triple Difference using All Tweets

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID attack	(5) Profanity	(6) Insult
Germany x After						
x Sensitive	-3.00***	-3.23***	-2.18***	-6.13***	-1.45**	-2.89***
	(1.12)	(0.96)	(0.76)	(1.81)	(0.68)	(0.76)
Germany x After	-0.04	-0.06	0.76*	0.30	-0.22	-0.20
	(0.32)	(0.43)	(0.42)	(0.54)	(0.31)	(0.47)
Sensitive x After	-0.54	0.14	-0.78	3.43**	0.92**	1.32**
	(0.91)	(0.74)	(0.48)	(1.63)	(0.41)	(0.61)
Germany x Sensitive	2.70*	1.86	2.25***	3.10	0.44	1.32
	(1.57)	(1.31)	(0.87)	(2.84)	(0.79)	(1.04)
Sensitive topic	9.72***	12.17***	6.27***	25.81***	2.57***	7.79***
	(1.40)	(1.15)	(0.61)	(2.73)	(0.54)	(0.93)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.129	0.178	0.099	0.249	0.095	0.164
Observations	2270652	2270652	2270652	2270652	2270652	2270652
Mean of Outcome	17.822	27.527	25.704	26.906	16.324	26.563
SD of Outcome	19.841	23.011	18.863	24.356	19.285	22.260

Notes. The table shows the main coefficients of the triple-difference estimations comparing the average hate intensities by users affected and unaffected by NetzDG and by sensitivity of the topic. Besides the treatment effects, all estimations control for country-specific events of regional/national elections and terrorist attacks, the day of the week the tweet was sent and an indicator if the tweet was sent at night. All estimations include a constant and year-month fixed effects, user fixed effects, and fixed effects for the account age in months when the respective tweet was posted. Standard errors are clustered at the user level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

with hate by hateful images or videos, we could expect an increase in the volumes of images and videos after NetzDG.

We analyse the effect of the law on other potential ways of hate expression estimating regressions which are similar to our baseline model with a set of tweet characteristics as dependent variables. Instead of continuous scores ranging from 0 to 100, we use tweeting style measures which are indicators of whether a tweet contains an image, video, any URL or a URL to the media from the top-25 media outlets in Germany. Additionally, we measure the change in the number of hashtags and words in a tweet and in the daily tweeting frequency. Table 2.9 suggests that, contrary to the hate speech intensity in tweets' texts, the tweeting style among German users did not change as compared to Austrian users. Our data, however, do not allow us to assess the content of images, videos, or links. Acknowledging our data limitations, we do not find any potential substitution patterns in tweeting due to the implementation of NetzDG, which could be overlooked by our hate speech measures.

Table 2.9: Substitution Patterns: Effect of NetzDG on Tweet Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Videos	Images	URL	Media Link	No. Hashtags	No. Words	Tweet. Freq.
Treated after T.	-0.00	0.00	0.03	0.02	0.13	1.20	-0.41
	(0.00)	(0.02)	(0.02)	(0.01)	(0.24)	(1.46)	(2.70)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.25	0.25	0.47	0.14	0.50	0.44	0.71
Observations	160161	160161	160161	160161	160161	160161	12002
Mean of Outcome	0.00	0.07	0.68	0.06	1.02	18.19	13.29
SD of Outcome	0.05	0.26	0.46	0.24	2.05	9.60	34.27

Notes. The table shows the main coefficients of the difference-in-differences estimations comparing tweet characteristics by users affected and unaffected by NetzDG. The columns contain different outcome types: Col (1)-(4) are indicators for a Video, Photos, URL, or Media Link in the tweet. Col. (5) and (6) are counts for the number of hashtags and words. Col. (7) analyzes the change in monthly tweeting frequency per user on a monthly basis. Besides the treatment effect, all estimations control for country-specific events of regional/national elections and terrorist attacks, the day of the week the tweet was sent and an indicator if the tweet was sent at night. All estimations include a constant and year-month fixed effects, user fixed effects, and fixed effects for the account age in months when the respective tweet was posted. Standard errors are clustered at the user level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

2.7 Potential Mechanisms of the Law

There are three potential mechanisms that can drive the reduction in average hate intensity scores:

1. Following NetzDG, platforms increase the removal of UGC containing hate;

2. Platform users decrease their expression of hate to avoid platform interventions;

3. Users with preferences for hate expression exit the large platforms subject to NetzDG or multihome, expressing hatred on platforms with weaker moderation.

The analysis in this section suggests that the main driver of the decrease in hate due to NetzDG in our setting is *self-censoring in the expression of hatred*. The effect of the regulation on the behaviour of the social network users is consistent with the findings of Huang, Hong, and Burtch 2016 that the integration of a social network into the review platform changes the volume and quality of UGC via shifts in user behaviour rather than in user composition. In the following, we investigate each of the potential mechanisms seperately.

2.7.1 Content Removal

To measure the extent of an increased content removal on Twitter, we would need to access the tweets that were taken down in Germany, which is not allowed by the platform. However, we retrieved the figures on user complaints and content removal that are officially provided by Twitter following the implementation of NetzDG, which are available semiannually. Figure 2.3 suggests that user complaints on hateful tweets due to NetzDG as well as subsequent removal did not increase until 2019, and the numbers of removed tweets were only increasing in the second half of the year 2019. At the same time, our baseline results (see Figure 2.6) suggest a decrease in hate across tweets already in quarters 2 and 4 of 2018, i.e. in the period when, according to the graph, the numbers of complaints and deleted tweets were not growing. Moreover, experimental evidence from the field suggests that in quarter 3 of 2020, when the removal numbers were very high, the platform deleted about 2.1% of hateful tweets expressing Holocaust denial and hate towards disabilities that were not reported to Twitter and 3.5% of hateful tweets that were reported (Jiménez Durán 2022). Due to such low scale, we suggest that content removal is not driving the results in our setting.

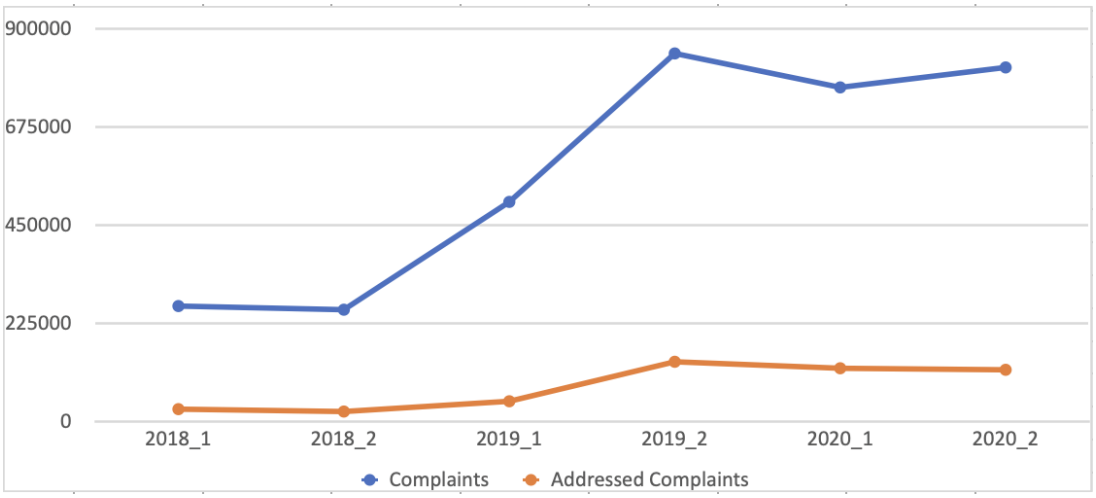


Figure 2.3: User Complaints and Tweet Deletions according to Twitter's NetzDG Reports

Note: Semiannual numbers of tweets reported by users as hateful according to NetzDG (in blue) and actually removed (in orange) according to Twitter NetzDG Reports.

Furthermore, we reran our baseline specification with tweets that were only posted until the end of 2018, i.e. until the deletions and user complaints according to Twitter's NetzDG reports did not experience an increase. Table 2.23 in the Appendix shows that also when deletions do not seem to play a major role, the hate intensities significantly decrease. Again, this suggests that content deletion is not the driving force of our results.

2.7.2 Self-Censorship

To better assess the likelihood of users self-censoring themselves in the expression of hate, we look at changes in the distribution of hate at the user level. Specifically, we run a similar analysis to the volume of hate at the user-month level, now focusing on the parameters of the distribution of hate intensity. Our series of regressions use as dependent variables the minimum, the median, and the maximum values of hate intensity for each hate measure at the user-month level. The results in Table 2.10 suggest that while there is no change in the monthly minimum values, there is a strong significant shift to the left in the median values of hate intensity of each user measured by severe toxicity, toxicity, identity attacks, and insults.

The maximum values also decrease for toxicity, identity attacks, and insult, but these effects are marginally significant. Additionally, when we consider the entire dataset with all tweets of our users (i.e, not only those tweets addressing sensitive topics), the hate intensity shifts to the right, with increases in the median and mean for the probability of identity attacks (these results are available upon request). This shows that while the entire German segment follows

the general trend on social networks towards an increase in mutual hate, the sensitive topics experience very significant and robust decrease in hate, i.e. "adjustment" in the language about these topics. *Importantly, if the shifts in the median of the hate intensity would have been due to tweets deletion by the platform, we would expect to measure stronger and more significant decreases in the maximum values, because the platform would focus on moderating tweets with the highest values of hate intensity.*

Table 2.10: Panel: The Changes in Hate Intentsity Distribution by User and Month

The Treatment Effect						
	(1)	(2)	(3)	(4)	(5)	(6)
	sev. Toxicity	Toxicity	Threat	ID Attack	Profanity	Insult
Min. Hate Score	0.46	0.13	0.60	0.38	-0.00	0.15
	(0.79)	(0.99)	(0.65)	(1.28)	(0.00)	(0.89)
Median Hate Score	-1.25*	-2.22***	-0.41	-2.48**	-0.00	-2.04***
	(0.74)	(0.74)	(0.75)	(1.01)	(0.00)	(0.70)
Max. Hate Score	-2.06	-1.96*	-3.13**	-2.05*	-0.02	-2.15*
	(1.32)	(1.03)	(1.52)	(1.15)	(0.02)	(1.10)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9546	9546	9546	9546	9546	9546
Groups	492	492	492	492	492	492

Notes. The table shows the coefficients of interest of the panel difference-in-difference estimations at the user-month level. The dependent variables (in rows) are the minimum, median, and maximum values of the corresponding hate intensity measures (in columns) computed at the user-month level. Besides the treatment effect, all estimations control for the country-specific share of tweets posted during night times and on days of regional/national elections and terrorist attacks. Standard errors are clustered at the user level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

We additionally compare the distribution of the average value of hate intensity measures before and after the introduction of NetzDG for the treated (Germany) and untreated (Austria) group suggests interesting patterns. The graphs in Figure 2.4 display the average of the mean hate score of each tweet. Figure 2.4 shows that the distribution of the hate score shifts towards higher scores in the middle part of the distribution in Austria after January 2018 and towards lower scores in Germany. Again, if the decreases in hate we measure were driven by the platform removing content, we would expect to observe a stronger change in the right tail of the distribution in Germany than in the middle part of the distribution. Hence, the treatment effect seems to be driven by users moderating their language in tweets about sensitive topics.

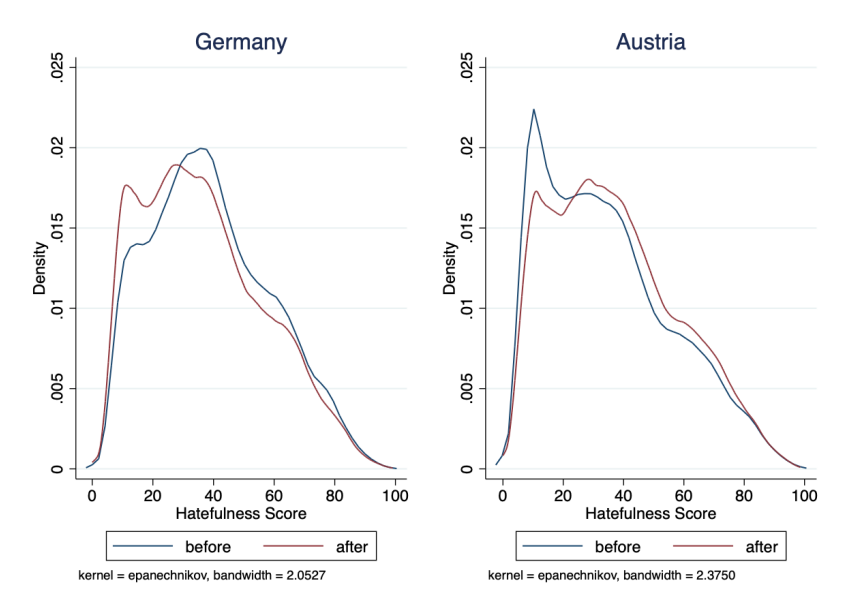


Figure 2.4: Distribution of Average Hate Intensity by Time Period and Treatment Status

Notes. These plots display the distribution of the hate intensity scores of each tweet by untreated (Austria) and treated (Germany) groups of users before and after NetzDG. Observations range from 1.5 years before to 1.5 years after NetzDG. The hate intensity scores are calculated as the averages of the scores of the six hate dimensions for each tweet.

The evidence we present in our study is consistent with user reactions to the revelations of the US government's privacy violations. Marthews and Tucker 2017 show that the Snowden revelations on mass digital surveillance had a chilling effect on online search such that after the revelations, users actively engaged in self-censorship by reducing their search volume for terms that could be perceived as sensitive.

2.7.3 User Migration

Anecdotal evidence suggests that users with strong preferences for uploading and viewing hateful content migrate to platforms with weaker or no content moderation in response to the efforts of large social media platforms such as Twitter and Facebook to moderate hateful content. A salient example of such migration behaviour is the messenger service Telegram with public channels, which was not subject to NetzDG until spring 2021. Due to the lax rules regarding any kind of UGC, Telegram attracts conspiracy theorists, right-wing extremists, and terrorists.[34] Telegram

[34] Spiegel Article:
`https://www.spiegel.de/international/world/the-telegram-billionaire-and-his-dark-empire-a-f27cb79f-86ae-48de-bdbd-8df604d07cc8`

reportedly received 25 million new users worldwide in a couple of days after the closure of Parler and media campaigns by Facebook and Twitter promising to increase their moderation efforts.[35]

These migration patterns might suggest relocation rather than mitigation of hate speech, although Rauchfleisch and Kaiser 2021 and Ali et al. 2021 conclude that deplatforming is still an effective tool to combat online hate speech due to the lower reach of hatred on smaller platforms. Moreover, as soon as smaller platforms grow in the number of users due to migration from the dominant platforms and reach the cutoff of 2 million users, they become subject to NetzDG and are forced to either moderate hate speech or leave the market.

We address the user migration by analyzing the user composition in our sample. Table 2.25 (in the Appendix) shows the share and the number of users in our sample i) who tweeted only before NetzDG was introduced, ii) only after NetzDG was introduced, and iii) who tweeted both before and after the introduction of NetzDG. About half of the users in our estimation sample (i.e., who tweeted about sensitive topics) was present on Twitter both before and after the introduction of NetzDG. Consistently with the general growth path of social media platforms (Hölig and Hasebrink 2020), more users joined than left our sample. This pattern is stronger in Germany than in Austria, and, surprisingly, the share of users leaving the sample is lower in Germany than in Austria.

Additionally, we replicate our baseline analysis using only tweets from users who tweeted in our sample at least twice before and after NetzDG. The results in Table 2.11 are confirming the baseline results and the effects are measured more precisely. Hence, our baseline results are driven by the users who continue tweeting on the platform after the implementation of NetzDG.

2.8 Conclusion

In recent years, social media platforms reportedly made high efforts to develop complex infrastructures and mechanisms to combat harmful content. Despite these initiatives, public concerns remained that large tech companies were not sufficiently addressing hate speech, considering that only a very small share of user complaints received responses. Hence, German legislators additionally imposed large platforms "to take on their responsibility in [the] question of deleting

[35] Politico Article: `https://www.politico.eu/article/telegram-far-right-extremist/`

Table 2.11: Robustness Check: Sample Restricted to Users Tweeting Before and After NetzDG

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID Attack	(5) Profanity	(6) Insult
Treated after T.	-1.86**	-2.24***	-0.57	-2.90***	-1.30**	-2.37***
	(0.77)	(0.60)	(0.77)	(0.91)	(0.53)	(0.60)
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes
Day of the Week	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.17	0.18	0.07	0.20	0.12	0.18
Observations	110612	110612	110612	110612	110612	110612
Mean of Outcome	29.09	41.56	34.37	56.00	19.39	35.80
SD of Outcome	23.82	22.41	24.78	27.70	19.75	21.78

Notes. The table replicates the baseline analysis, but for the subset of users who are observed at least twice before and after NetzDG came into effect. It shows the main coefficients of the difference-in-differences estimations comparing the hate intensity in tweets by those staying users that are affected and unaffected by the law (NetzDG). The columns contain the outcome measures discussed in the data section: Continuous scores ranging from 0 to 100 with regard to severe toxicity, toxicity etc. as calculated by Perspective API. The coefficient *Treated after T.* shows the change in hate intensity in terms of percentage points for users located in Germany after NetzDG became effective. Besides the treatment effect, all estimations control for country-specific events of regional/national elections and terrorist attacks, the day of the week the tweet was sent and an indicator if the tweet was sent at night. All estimations include a constant and year-month fixed effects, user fixed effects, and fixed effects for the account age in months when the respective tweet was posted. Standard errors are clustered at the user level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$.

criminal content" (Heiko Maas, federal minister for justice and consumer protection)[36] and implemented NetzDG, a law requiring large social media platforms to timely remove user-reported hateful content. The critiques of the law, including politicians and members of civil society, mention the threats of over-policing of digital communication and restrictions on freedom of speech. Furthermore, opponents claimed that platforms might not comply due to the law's perceived vagueness, suggesting it could increase legal uncertainty. Our paper adds to this debate, suggesting that the implementation of NetzDG led to a decrease in the intensity and volume of hate within German tweets about sensitive topics.

Although Twitter claims that the NetzDG did not affect content moderation as most of hate speech is removed due to its internal governance policy (Liesching et al. 2021), we find that legal regulation can contribute to curbing harmful content even when platforms already have governance rules for the same purpose. While the platform's governance rules apply to tweets by

[36]Politico Article: `https://www.politico.eu/article/germany-unveils-law-with-big-fines-for-hate-speech-on-social-media/`

users located in both Austria and Germany, our results suggest an additional reduction in hate speech posted by users located in Germany. We find robust effects of the regulation on decreasing the intensity of (severe) toxicity, profanity, insults, and identity attacks in Germany as opposed to Austria by about 5-6% of the mean. Moreover, the volume of potentially unlawful tweets decreased by 11% in the number of original hateful tweets. Additionally, we find that hateful tweets generally receive higher user engagement. Hence, the reduction in the number of original hateful tweets decreases the exposure to hate even more due to prevented impressions and user engagement with hateful content. In line with the law's objective, we show that the decrease in hate intensity is driven by tweets addressing sensitive topics of migration and religion, while the rest of tweets and other tweeting patterns are unaffected by the regulation.

Our analyses address the three potential mechanisms driving the decrease in hate intensity due to NetzDG. Although data limitations do not allow us to fully rule out that platforms delete more hateful UGC, we show that our results are mostly driven by the decrease in the median level of hate intensity in sensitive tweets, while the maximum level of hate intensity was not significantly affected. This suggests that platform users self-censor in the expression of hatred following the implementation of NetzDG.

The implementation of NetzDG has inspired many countries to design similar national regulations. Later on, German NetzDG served as a blueprint for the EU-wide Digital Services Act, which has recently come into force. Importantly, we measure these effects before the recent event that changed the face of the platform. In the fall of 2022, after the acquisition of Twitter by Elon Musk, media reported lay-offs of many thousands of content moderators who monitored the prevalence of abusive content and misinformation on the platform.[37] This was an alarming event for the public and the expert community, and while lay-offs affected many countries, the German Twittersphere was deemed one of the most legally protected due to NetzDG. Two months later, Twitter faced a lawsuit in Germany for failing to timely remove illegal content.[38] The recent developments highlight that regulation is of paramount importance for protecting individuals from offline (psychological) harm caused by online presence.

[37] Deutsche Welle Article: `https://www.dw.com/en/twitters-sacking-of-content-moderators-will-backfire-experts-warn/a-63778330`

[38] Euractive Article: `https://www.euractiv.com/section/platforms/news/twitter-faces-lawsuit-in-germany-for-failure-to-remove-anti-semitic-content/`

Hence, our findings are crucial in this context as they offer insights for policy makers, platform managers and the general public about the potential implications and mechanisms of legally imposed content moderation.

Appendices

A Further Information on Data

Table 2.12: Outcome Variables for an Exemplary Tweet as Computed by Perspective API

Example tweet, translated to English:
"We have pulled the teeth out of pagan + witch-killing Christianity... Islam is waiting"

Outcome	**Score**	**Definition**
Severe Toxicity	58.09524	A very hateful, aggressive, disrespectful comment or otherwise very likely to make users leave a discussion or give up on sharing their perspective. This attribute is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words.
Toxicity	81.43812	A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion.
Threat	65.58015	Describes an intention to inflict pain, injury, or violence against an individual or group.
Identity Attack	91.92697	Negative or hateful comments targeting someone because of their identity.
Profanity	32.70008	Swear words, curse words, or other obscene or profane language.
Insult	65.19685	Insulting, inflammatory, or negative comment towards a person or a group of people.

Notes. This table shows the estimated hate intensity scores with regard to all hate dimensions used in this analysis. The last column includes the definitions of the dimensions as defined by Perspective.

Table 2.13: Original Example Tweets in our Sample

Outcome	Value	Example Tweet
Severe Toxicity	1	eckelhafter drecksack...dann verpisst euch hier,ihr hurensöhne,fuck islam
Toxicity	0.99	wie dumm bist du eigentlich? bei dir ist gleich jeder ein pkkler terrorist.du gehörst zurück gepudert und abgetrieben.
Threat	0.99	diesem typ wünsche ich den tod durch einen dieser krimigranten.
Identity Attack	1	jepp, katholiken ficken kinder, moslems schlagen ihnen die fresse ein und schneiden mädchen die klitoris ab. juden und moslems lassen tiere liebevoll ausbluten. religion ist ein hurensohn.
Profanity	0.99	dieses arschkriechen vor dem scheiß islam ist echt nur noch zum kotzen
Insult	0.99	[...] diese deppen kapieren nie wie völkisch moslems sind

Table 2.14: Translated Example Tweets in our Sample

Outcome	Value	Example Tweet
Severe Toxicity	1	disgusting scumbag...then fuck off here, you sons of bitches, fuck islam
Toxicity	0.99	how stupid are you? for you every pkkler is a terrorist. you belong back powdered and aborted.
Threat	0.99	i wish this guy death by one of these criminals.
Identity Attack	1	yeah, catholics fuck children, muslims smash their faces and cut off girls' clitorises. jews and muslims lovingly bleed animals. religion is a son of a bitch.
Profanity	0.99	this ass-kissing of the fucking islam is really just to vomit
Insult	0.99	[...] these morons never get how nationalistic muslims are

Table 2.15: Summary Table of Tweet Characteristics

	Germany					Austria				
Variable	Mean	P50	SD	Min	Max	Mean	P50	SD	Min	Max
Severe Toxicity	31.05	29	23.69	0.00	100.00	28.32	18	24.70	0.00	100.00
Toxicity	43.68	46	21.74	0.00	100.00	41.47	42	23.47	0.00	100.00
Threat	36.21	22	25.45	0.00	99.93	32.47	20	23.86	0.00	99.46
Identity Attack	58.61	63	26.58	0.00	100.00	55.58	60	28.98	0.00	100.00
Profanity	20.51	11	20.12	0.00	100.00	20.42	11	20.99	0.00	99.95
Insult	37.54	36	21.42	0.00	99.72	36.78	36	23.01	0.00	99.72
No. of Retweets	4.89	0	22.17	0.00	911.00	2.62	0	13.84	0.00	779.00
No. of Likes	8.28	0	41.28	0.00	1398.00	5.12	0	26.19	0.00	1711.00
No. of Replies	1.28	0	6.71	0.00	292.00	0.86	0	3.30	0.00	275.00
Video in tweet	0.00	0	0.04	0.00	1.00	0.00	0	0.06	0.00	1.00
Photo	0.08	0	0.28	0.00	1.00	0.05	0	0.23	0.00	1.00
URL	0.74	1	0.44	0.00	1.00	0.59	1	0.49	0.00	1.00
Link to media outlet	0.08	0	0.27	0.00	1.00	0.03	0	0.18	0.00	1.00
No. of Words	18.33	15	9.63	1.00	57.00	17.96	15	9.56	1.00	52.00
Tweeted at night	0.08	0	0.28	0.00	1.00	0.06	0	0.24	0.00	1.00
Terrorist attack in country	0.02	0	0.15	0.00	1.00	0.00	0	0.06	0.00	1.00
Election in country	0.01	0	0.11	0.00	1.00	0.01	0	0.08	0.00	1.00

Sample: All tweets including a migration or religion specific buzzword.

Table 2.16: Raw Pairwise Correlations among Outcome Variables

	Severe Toxicity	Toxicity	Threat	Identity Attack	Profanity	Insult
Severe Toxicity	1.00					
Toxicity	0.90***	1.00				
Threat	0.57***	0.53***	1.00			
Identity Attack	0.75***	0.85***	0.38***	1.00		
Profanity	0.86***	0.81***	0.45***	0.59***	1.00	
Insult	0.85***	0.93***	0.39***	0.80***	0.86***	1.00
Observations	160474					

Notes: The sample includes all tweets with the filtering buzzwords for migration and religion.

(a) Severe Toxicity

(b) Toxicity

(c) Threat

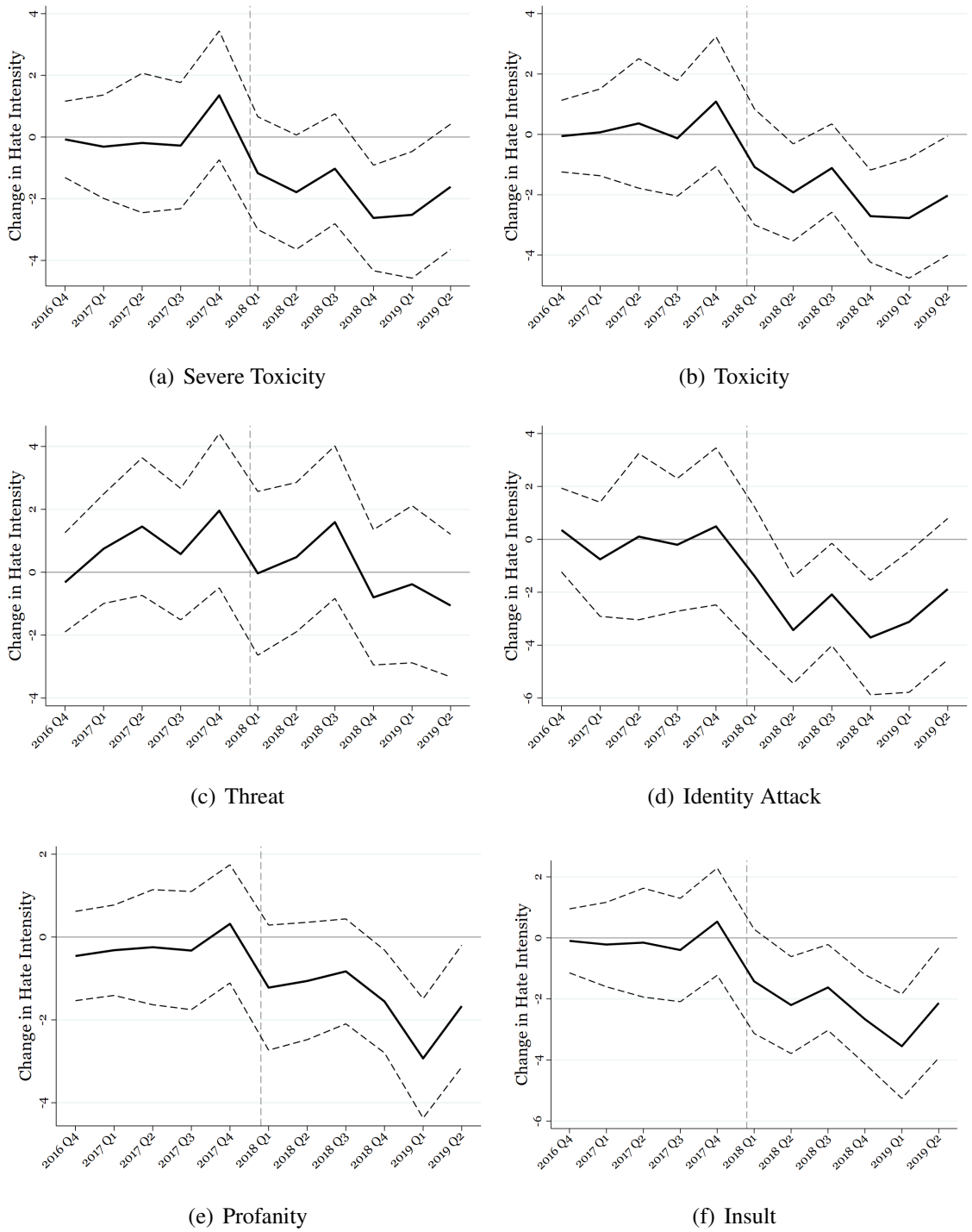
(d) Identity Attack

(e) Profanity

(f) Insult

Figure 2.5: Distributions of the Hate Scores by Perspective API and four Human Classifiers

B Further Analyses and Robustness Checks



(a) Severe Toxicity (b) Toxicity

(c) Threat (d) Identity Attack

(e) Profanity (f) Insult

Figure 2.6: Quarterly Treatment Effects with Pre-trends

Notes. The plot shows the treatment effects for Q4 2016 - Q2 2019 for the six hate dimensions. Shown is the coefficient of the interaction of a treated tweet (posted by a user located in Germany) with different timings for NetzDG and the 90% confidence interval, while controlling for country specific events of regional/national elections and terrorist attacks, the day of the week the tweet was sent and an indicator if the tweet was sent at night. All estimations include a constant and user fixed effects, year-month fixed effects, and fixed effects for the account age in months when the respective tweet was posted. Standard errors are clustered at the user level. The vertical line indicates the date NetzDG became effective.

Table 2.17: The Effect of NetzDG on the Intensity of Hate in Tweets (OLS with FE, All Coefficients)

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID Attack	(5) Profanity	(6) Insult
Treated after T.	-1.89***	-2.15***	-0.76	-2.63***	-1.33***	-2.18***
	(0.68)	(0.54)	(0.71)	(0.81)	(0.50)	(0.55)
Tweeted at night	0.72	0.71	0.13	0.08	0.87	0.74
	(0.66)	(0.61)	(0.40)	(0.75)	(0.59)	(0.63)
Tuesday	-0.31	-0.13	-0.24	-0.22	-0.32*	-0.06
	(0.20)	(0.19)	(0.24)	(0.24)	(0.18)	(0.19)
Wednesday	-0.55**	-0.39*	-0.69***	-0.72***	-0.42**	-0.32
	(0.22)	(0.22)	(0.26)	(0.25)	(0.20)	(0.21)
Thursday	-0.53**	-0.54**	-0.81***	-0.56**	-0.41**	-0.29
	(0.22)	(0.21)	(0.25)	(0.24)	(0.20)	(0.22)
Friday	-0.08	0.07	0.09	-0.48*	-0.11	0.03
	(0.25)	(0.25)	(0.23)	(0.29)	(0.22)	(0.22)
Saturday	0.11	0.41	-0.38	0.36	0.12	0.50*
	(0.31)	(0.28)	(0.31)	(0.33)	(0.28)	(0.29)
Sunday	0.03	0.10	-0.21	0.21	-0.07	0.01
	(0.23)	(0.21)	(0.29)	(0.26)	(0.19)	(0.21)
Terrorist attack in country	0.25	0.28	1.36*	-0.14	-0.10	-0.18
	(0.50)	(0.51)	(0.70)	(0.65)	(0.40)	(0.47)
Election in country	-0.32	-0.68	-0.55	-0.33	-0.59	-0.27
	(0.56)	(0.55)	(0.71)	(0.63)	(0.50)	(0.53)
Constant	30.70***	43.51***	35.27***	58.45***	21.01***	37.90***
	(0.23)	(0.21)	(0.28)	(0.29)	(0.21)	(0.20)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.163	0.177	0.072	0.199	0.123	0.178
Observations	160165	160165	160165	160165	160165	160165
Mean of Outcome	29.973	42.810	34.735	57.419	20.476	37.242
SD of Outcome	24.126	22.463	24.900	27.590	20.468	22.059

Clustered standard errors in parentheses, clustered at the user level. All models include an intercept.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 2.18: The Effect of NetzDG on the Intensity of Hate in Tweets (OLS without FE, All Coefficients)

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID Attack	(5) Profanity	(6) Insult
Germany	3.23	3.04*	3.82***	3.65*	0.51	1.49
	(1.99)	(1.56)	(1.39)	(1.99)	(1.32)	(1.57)
Treated after T.=1	-2.18*	-3.26***	-0.87	-4.69***	-1.70*	-3.14***
	(1.31)	(1.12)	(1.00)	(1.59)	(0.91)	(1.17)
Tweeted at night	2.75***	2.83***	0.31	3.28***	2.44***	3.11***
	(0.88)	(0.83)	(0.56)	(1.15)	(0.76)	(0.94)
verified=1	-5.26***	-4.45***	-3.19***	-3.54**	-4.60***	-4.56***
	(1.47)	(1.36)	(0.82)	(1.79)	(1.08)	(1.52)
No. of Followers	-0.00**	-0.00***	-0.00***	-0.00***	-0.00***	-0.00***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Tuesday	-0.20	-0.02	-0.16	-0.06	-0.20	0.06
	(0.23)	(0.22)	(0.25)	(0.29)	(0.21)	(0.23)
Wednesday	-0.68***	-0.49**	-0.77***	-0.77***	-0.51**	-0.38
	(0.23)	(0.23)	(0.27)	(0.28)	(0.21)	(0.24)
Thursday	-0.53**	-0.53**	-0.85***	-0.46*	-0.41*	-0.26
	(0.26)	(0.23)	(0.27)	(0.28)	(0.22)	(0.24)
Friday	0.08	0.17	0.20	-0.42	-0.02	0.12
	(0.27)	(0.27)	(0.23)	(0.32)	(0.23)	(0.25)
Saturday	0.23	0.62**	-0.49	0.59	0.24	0.72**
	(0.37)	(0.31)	(0.34)	(0.36)	(0.31)	(0.32)
Sunday	0.13	0.31	-0.39	0.47	0.08	0.29
	(0.27)	(0.24)	(0.32)	(0.31)	(0.22)	(0.25)
Terrorist attack in country	0.23	0.28	1.37**	-0.25	0.01	-0.07
	(0.57)	(0.54)	(0.69)	(0.71)	(0.42)	(0.49)
Election in country	-0.45	-1.01	-0.40	-1.02	-0.72	-0.64
	(0.67)	(0.67)	(0.76)	(0.81)	(0.57)	(0.64)
Constant	38.70***	49.59***	45.28***	60.64***	26.12***	41.80***
	(3.46)	(3.02)	(2.51)	(4.37)	(2.59)	(3.02)
month indicators	Yes	Yes	Yes	Yes	Yes	Yes
account age	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.05	0.07	0.02	0.08	0.04	0.06
Observations	160407	160407	160407	160407	160407	160407
Mean of Outcome	29.97	42.81	34.73	57.41	20.48	37.24
SD of Outcome	24.13	22.47	24.90	27.59	20.47	22.06

Clustered standard errors in parentheses, clustered at the user level. * p<0.10, ** p<0.05, *** p<0.01.

All models include an intercept.

Table 2.19: Robustness Check: Sample without Users Living Outside Germany/Austria

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID Attack	(5) Profanity	(6) Insult
Treated after T.	-2.14***	-2.45***	-0.54	-3.27***	-1.25**	-2.36***
	(0.75)	(0.58)	(0.72)	(0.87)	(0.53)	(0.58)
Constant	29.80***	42.81***	34.77***	57.55***	20.14***	37.15***
	(0.29)	(0.22)	(0.25)	(0.32)	(0.18)	(0.20)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes
Day of the Week	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.14	0.17	0.07	0.19	0.11	0.17
Observations	140872	140872	140872	140872	140872	140872
Mean of Outcome	29.10	42.00	34.60	56.39	19.75	36.37
SD of Outcome	23.52	22.24	24.86	27.56	19.84	21.71

Standard errors in parentheses, clustered at the user level. All models include an intercept.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 2.20: Robustness Check: Baseline Analysis Excluding Transition Period (July'17-Dec'17)

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID Attack	(5) Profanity	(6) Insult
Treated after T.	-1.72**	-2.11***	-0.51	-2.76***	-1.27**	-2.20***
	(0.76)	(0.57)	(0.81)	(0.82)	(0.55)	(0.58)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.16	0.18	0.07	0.20	0.12	0.18
Observations	135883	135883	135883	135883	135883	135883
Mean of Outcome	29.69	42.64	34.63	57.28	20.41	37.18
SD of Outcome	23.99	22.41	24.84	27.56	20.45	22.02

Standard errors in parentheses, clustered at the user level. All models include an intercept.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 2.21: Robustness Check: Setting NetzDG to January 2017

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID Attack	(5) Profanity	(6) Insult
Treated before T.	-0.80	-0.75	0.75	-1.53	-0.60	-1.09
	(0.77)	(0.73)	(0.95)	(0.97)	(0.56)	(0.71)
Constant	30.31***	43.13***	34.34***	58.17***	20.71***	37.73***
	(0.37)	(0.35)	(0.46)	(0.45)	(0.26)	(0.33)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes
Day of the Week	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.16	0.18	0.07	0.20	0.12	0.18
Observations	160161	160161	160161	160161	160161	160161
Mean of Outcome	29.97	42.81	34.73	57.42	20.48	37.24
SD of Outcome	24.13	22.46	24.90	27.59	20.47	22.06

Clustered standard errors in parentheses, clustered at the user level. All models include an intercept.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 2.22: OLS with FE using all tweets

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID attack	(5) Profanity	(6) Insult
Treated after T.	-0.19	-0.15	0.61	0.35	-0.27	-0.27
	(0.37)	(0.48)	(0.44)	(0.60)	(0.31)	(0.47)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.113	0.161	0.092	0.171	0.094	0.156
Observations	2270652	2270652	2270652	2270652	2270652	2270652
Mean of Outcome	17.822	27.527	25.704	26.906	16.324	26.563
SD of Outcome	19.841	23.011	18.863	24.356	19.285	22.260

Clustered standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 2.23: Triple Difference using All Tweets Until 2018

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID attack	(5) Profanity	(6) Insult
Germany x After x Sensitive	-3.12**	-3.23***	-2.29***	-6.35***	-1.27*	-2.76***
	(1.23)	(1.13)	(0.79)	(2.20)	(0.75)	(0.89)
Germany x After	0.20	0.26	1.19***	0.57	-0.08	0.02
	(0.29)	(0.39)	(0.39)	(0.47)	(0.28)	(0.42)
Sensitive x After	-0.39	0.05	-0.99*	3.38	0.84	1.22
	(1.08)	(0.97)	(0.54)	(2.06)	(0.51)	(0.78)
Germany x Sensitive	2.76*	1.89	2.38***	3.15	0.42	1.29
	(1.60)	(1.35)	(0.88)	(2.85)	(0.81)	(1.06)
Sensitive topic	9.61***	12.09***	6.15***	25.74***	2.55***	7.80***
	(1.43)	(1.19)	(0.62)	(2.74)	(0.57)	(0.95)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.137	0.189	0.108	0.261	0.098	0.171
Observations	1796042	1796042	1796042	1796042	1796042	1796042
Mean of Outcome	17.920	27.445	25.809	27.010	16.191	26.350
SD of Outcome	19.888	22.977	19.022	24.470	19.144	22.128

Notes. The table replicates the Triple Difference analysis, but only consideres observations until the end of the year 2018. Clustered standard errors in parentheses, clustered at the user level. All models include an intercept.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 2.24: Panel: Volume of Outcome Variables by User-Month

	Volume Hateful			Volume Total
	(1) Sev. Toxicity	(2) Toxicity	(3) ID Attack	(4)
Treated after T.	-0.09**	-0.06*	-0.12*	-0.04
	(0.04)	(0.04)	(0.06)	(0.07)
month indicators	Yes	Yes	Yes	Yes
R^2	0.016	0.010	0.020	0.036
Observations	6292	6292	6292	6292
Mean of Outcome	0.261	0.188	0.795	1.913
SD of Outcome	0.539	0.437	0.967	1.172

Robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

All models include an intercept.

Table 2.25: User Composition

	Germany		Austria		Total	
	Share	Count	Share	Count	Share	Count
Stayed in Sample	0.46	272	0.47	351	0.47	623
Joined Sample	0.34	205	0.29	215	0.31	420
Left Sample	0.20	120	0.24	174	0.22	294
Observations		597		740		1337

Notes. The table shows the share and the absolute number of users observed in either both sample periods (before and after NetzDG) or only before or only after NetzDG.

C User Engagement

Table 2.26: User Engagement with Potentially Unlawful Tweets - Log Retweets 1-3

	(1)	(2)	(3)
	sev. Toxicity	Toxicity	Threat
Germany × AfterT	0.09	0.09	0.09
	(0.06)	(0.06)	(0.05)
Severely toxic	0.07***		
	(0.02)		
Germany × Severely toxic	-0.03		
	(0.02)		
AfterT × Severely toxic	-0.03		
	(0.02)		
Germany × AfterT × Severely toxic	0.04		
	(0.03)		
Toxic		0.04***	
		(0.02)	
Germany × Toxic		-0.02	
		(0.02)	
AfterT × Toxic		-0.05*	
		(0.03)	
Germany × AfterT × Toxic		0.05	
		(0.04)	
Threat			0.06***
			(0.02)
Germany × Threat			-0.01
			(0.02)
AfterT × Threat			0.03
			(0.02)
Germany × AfterT × Threat			0.04
			(0.03)
Month FE	Yes	Yes	Yes
User FE	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes
Day of the Week	Yes	Yes	Yes
R^2	0.54	0.54	0.54
Observations	160161	160161	160161
Mean of Outcome	0.56	0.56	0.56
SD of Outcome	1.00	1.00	1.00

Clustered standard errors in parentheses, clustered at the user level. All models include an intercept.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 2.27: User Engagement with Potentially Unlawful Tweets - Log Retweets 4-6

	(1)	(2)	(3)
	ID Attack	Profanity	Insult
Germany × AfterT	0.08	0.09	0.09
	(0.06)	(0.06)	(0.06)
ID Attack	0.05***		
	(0.01)		
Germany × ID Attack	-0.02		
	(0.02)		
AfterT × ID Attack	-0.01		
	(0.02)		
Germany × AfterT × ID Attack	0.05*		
	(0.03)		
Profanity		0.04	
		(0.02)	
Germany × Profanity		-0.02	
		(0.03)	
AfterT × Profanity		-0.03	
		(0.04)	
Germany × AfterT × Profanity		0.05	
		(0.05)	
Insult			0.04*
			(0.02)
Germany × Insult			-0.03
			(0.03)
AfterT × Insult			-0.05*
			(0.03)
Germany × AfterT × Insult			0.08*
			(0.05)
Month FE	Yes	Yes	Yes
User FE	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes
Day of the Week	Yes	Yes	Yes
R^2	0.54	0.54	0.54
Observations	160161	160161	160161
Mean of Outcome	0.56	0.56	0.56
SD of Outcome	1.00	1.00	1.00

Clustered standard errors in parentheses, clustered at the user level. All models include an intercept.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 2.28: User Engagement with Potentially Unlawful Tweets - Log Likes 1-3

	(1) sev. Toxicity	(2) Toxicity	(3) Threat
Germany × AfterT	0.00	0.00	0.01
	(0.07)	(0.07)	(0.07)
Severely toxic	0.10***		
	(0.02)		
Germany × Severely toxic	-0.03		
	(0.03)		
AfterT × Severely toxic	-0.07**		
	(0.04)		
Germany × AfterT × Severely toxic	0.07		
	(0.05)		
Toxic		0.09***	
		(0.02)	
Germany × Toxic		-0.01	
		(0.03)	
AfterT × Toxic		-0.07	
		(0.04)	
Germany × AfterT × Toxic		0.05	
		(0.06)	
Threat			0.00
			(0.03)
Germany × Threat			0.01
			(0.03)
AfterT × Threat			0.04
			(0.03)
Germany × AfterT × Threat			-0.01
			(0.03)
Month FE	Yes	Yes	Yes
User FE	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes
Day of the Week	Yes	Yes	Yes
R^2	0.55	0.55	0.55
Observations	160161	160161	160161
Mean of Outcome	0.75	0.75	0.75
SD of Outcome	1.14	1.14	1.14

Clustered standard errors in parentheses, clustered at the user level. All models include an intercept.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 2.29: User Engagement with Potentially Unlawful Tweets - Log Likes 4-6

	(1) ID Attack	(2) Profanity	(3) Insult
Germany × AfterT	-0.01	0.01	0.00
	(0.07)	(0.07)	(0.07)
ID Attack	0.08***		
	(0.02)		
Germany × ID Attack	-0.02		
	(0.02)		
AfterT × ID Attack	-0.01		
	(0.03)		
Germany × AfterT × ID Attack	0.08**		
	(0.03)		
Profanity		0.07**	
		(0.03)	
Germany × Profanity		0.00	
		(0.04)	
AfterT × Profanity		-0.02	
		(0.05)	
Germany × AfterT × Profanity		0.05	
		(0.07)	
Insult			0.10***
			(0.02)
Germany × Insult			-0.01
			(0.03)
AfterT × Insult			-0.08*
			(0.04)
Germany × AfterT × Insult			0.10
			(0.06)
Month FE	Yes	Yes	Yes
User FE	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes
Day of the Week	Yes	Yes	Yes
R^2	0.55	0.55	0.55
Observations	160161	160161	160161
Mean of Outcome	0.75	0.75	0.75
SD of Outcome	1.14	1.14	1.14

Clustered standard errors in parentheses, clustered at the user level. All models include an intercept.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 2.30: User Engagement with Potentially Unlawful Tweets - Log Replies 1-3

	(1)	(2)	(3)
	sev. Toxicity	Toxicity	Threat
Germany × AfterT	-0.02	-0.02	-0.02
	(0.05)	(0.05)	(0.05)
Severely toxic	0.03**		
	(0.02)		
Germany × Severely toxic	0.00		
	(0.02)		
AfterT × Severely toxic	-0.02		
	(0.02)		
Germany × AfterT × Severely toxic	-0.02		
	(0.03)		
Toxic		0.05***	
		(0.01)	
Germany × Toxic		-0.02	
		(0.02)	
AfterT × Toxic		-0.05**	
		(0.02)	
Germany × AfterT × Toxic		0.02	
		(0.04)	
Threat			0.01
			(0.01)
Germany × Threat			0.00
			(0.01)
AfterT × Threat			0.01
			(0.01)
Germany × AfterT × Threat			-0.01
			(0.02)
Month FE	Yes	Yes	Yes
User FE	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes
Day of the Week	Yes	Yes	Yes
R^2	0.43	0.43	0.43
Observations	160161	160161	160161
Mean of Outcome	0.32	0.32	0.32
SD of Outcome	0.65	0.65	0.65

Clustered standard errors in parentheses, clustered at the user level. All models include an intercept.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 2.31: User Engagement with Potentially Unlawful Tweets - Log Replies 4-6

	(1)	(2)	(3)
	ID attack	Profanity	Insult
Germany × AfterT	-0.03	-0.02	-0.02
	(0.05)	(0.05)	(0.04)
ID Attack	0.05***		
	(0.02)		
Germany × ID Attack	-0.03		
	(0.02)		
AfterT × ID Attack	-0.03		
	(0.02)		
Germany × AfterT			
× ID Attack	0.03		
	(0.02)		
Profanity		0.03	
		(0.02)	
Germany × Profanity		0.01	
		(0.02)	
AfterT × Profanity		-0.00	
		(0.03)	
Germany × AfterT × Profanity		-0.01	
		(0.04)	
Insult			0.04**
			(0.02)
Germany × Insult			-0.01
			(0.03)
AfterT × Insult			-0.03
			(0.03)
Germany × AfterT × Insult			0.02
			(0.04)
Month FE	Yes	Yes	Yes
User FE	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes
Day of the Week	Yes	Yes	Yes
R^2	0.43	0.43	0.43
Observations	160161	160161	160161
Mean of Outcome	0.32	0.32	0.32
SD of Outcome	0.65	0.65	0.65

Clustered standard errors in parentheses, clustered at the user level. All models include an intercept.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Chapter 3

YouTube “Adpocalypse”: The YouTubers’ Journey from Ad-Based to Patron-Based Revenues

Published as

Andres, Raphaela, Michelangelo Rossi and Mark Tremblay (2023), YouTube “Adpocalypse”: The YouTubers’ Journey from Ad-Based to Patron-Based Revenues, ZEW Discussion Paper No. 23-059, Mannheim.

3.1 Introduction

The Creator Economy has experienced an unparalleled increase in growth, fundamentally altering how individuals, ranging from artists to influencers, build careers and achieve financial autonomy. Over the past few decades, this dynamic ecosystem has flourished, and forecasts suggest it will continue to drive a multi-billion-dollar industry (Florida 2022; El Sanyoura and Anderson 2022). At the core of this transformation are content creators (CC), whose contributions are fundamental to the platform business model. Unlike traditional media industries, where content creation was centralized, modern digital platforms entrust this task to an array of independent CCs (Bhargava 2022). One of the first platforms pioneering content creation at scale is YouTube, which currently boasts more than 100 million channels and over 500 hours of content uploaded each minute.[39] Platforms like YouTube, TikTok, or Twitch have thrived on a multi-sided business model (Cusumano, Gawer, and Yoffie 2019), capitalizing on the cross-network effects of three primary stakeholders: CCs, users, and advertisers. CCs craft content that captivates users who not only consume this content but also engage with the advertising embedded in it. Advertisers, in turn, benefit from the connections with users and compensate the platform, which then rewards the creators. However, the viability of this model depends on the satisfaction of multiple actors and the proper matching of users, content, and advertising.

To meet the challenges of this type of business model, some online platforms have recently explored ways to encourage consumers to pay directly for content (thus reducing dependence on the advertising side) or to have greater control over the content produced by third-parties in order to properly match their content with advertising and user needs. In the latter case, the platform assumes the role of overseeing and curating the content produced by creators. This intricate interplay between enabling CCs and controlling the ecosystem underscores a critical trade-off within the Creator Economy as it has been described by Boudreau and Hagiu 2009, Hagiu and Wright 2015, and Hagiu and Wright 2019.

These important connections across users suggests that platform strategies for innovating and adjusting multi-sided business models cannot operate in isolation. CCs face low barriers to entry on these platforms, allowing them to switch between or establish a presence on multiple platforms simultaneously. Hence, when a platform aims to explore changes in its business model, such as increasing or decreasing content moderation or curation, it is critical for platforms to

[39] `https://blog.YouTube/press`

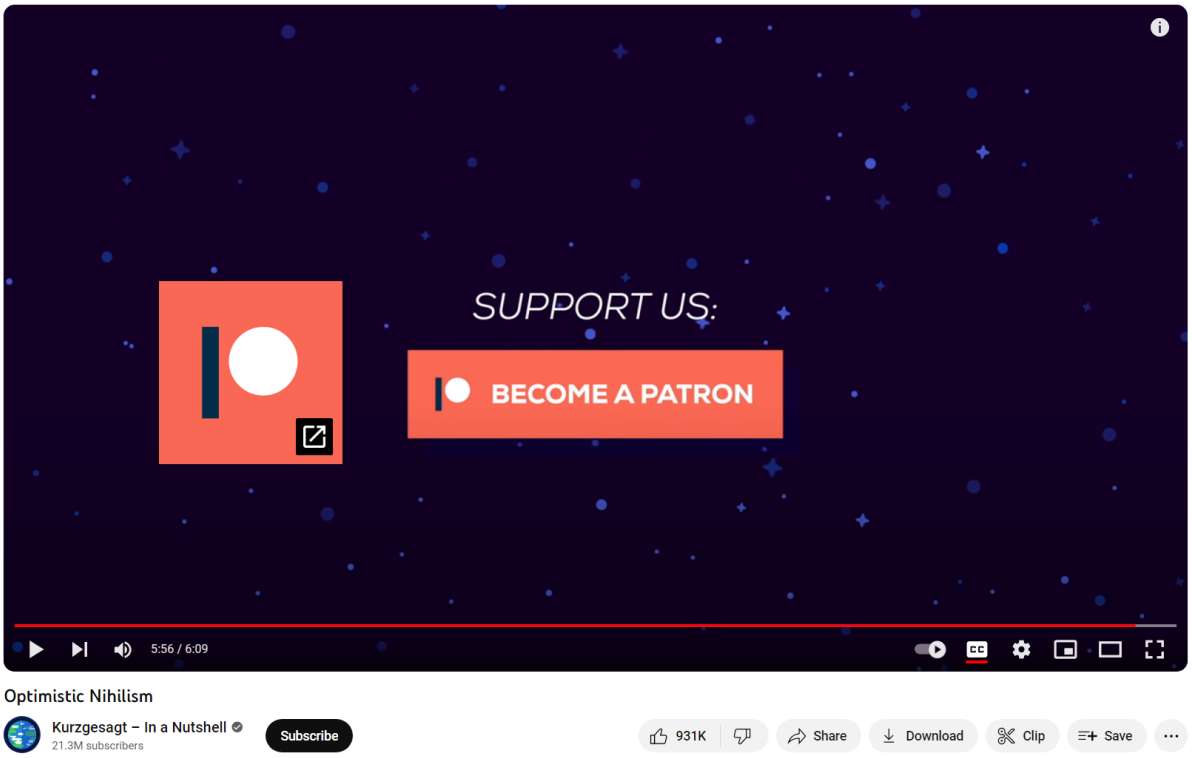
analyze the response of CCs. This includes their potential migration to another platform or increased efforts in creating content on competing platforms.

In this paper, we offer empirical evidence of how CCs and users react to changes in a platform's moderation policies. Our focus centers on two of the most successful platforms for CCs: YouTube and Patreon. YouTube primarily monetizes content through advertising revenue. Creators can enable ads on their videos and earn a share of the revenue generated from ad views and interactions. Patreon is a platform designed to help creators generate income directly from their supporters or "patrons" through monthly subscriptions. Creators on Patreon offer exclusive content to their subscribers in exchange for recurring monthly payments. Patreon primarily serves as a "membership portal" and does not actively direct users to discover other creators.[40] As a result, CCs are often required to establish a presence on other platforms, such as YouTube or Facebook, in order to attract users who might otherwise remain unaware of their content. As depicted in Figure 3.1, many CCs on YouTube actively guide their audience to their Patreon page by explicitly mentioning it within their videos or in the video description. In addition, Patreon's content moderation policies are much more permissive and do not have the imperative to accommodate the needs of advertisers by embedding content that aligns with their values, allowing for a more creator-centric environment.

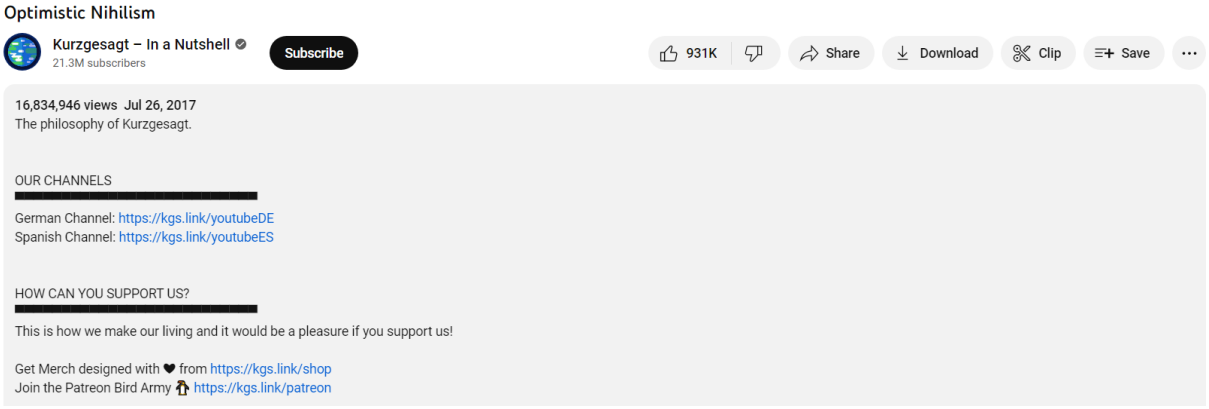
To determine how content moderation impacts content creator efforts across platforms that offer different revenue models, we take advantage of one of the most notable advertising boycotts in YouTube's history: the YouTube "Adpocalypse". In 2017, several institutional advertisers discovered that their YouTube ads were appearing alongside content that was inappropriate or extremist. This realization led to a mass exodus of major advertisers, resulting in a significant drop in advertising revenue.[41] In response to this advertising boycott, YouTube introduced stricter content moderation policies in January 2018. These changes aimed to improve the platform's ad-friendly environment by implementing stricter review and demonetization procedures for CCs who violated the new guidelines. Our analysis focuses on CCs who are active on Patreon, and we examine changes in content provision before and after the YouTube "Adpocalypse" between

[40] To better understand how the platform relies on users knowing the exact name of the creator they want to support, please see Patreon Support: `https://support.patreon.com/hc/en-us/articles/360028768352-Find-creators-on-Patreon`

[41] See, for example, The Guardian: `https://www.theguardian.com/technology/2017/mar/25/google-youtube-advertising-extremist-content-att-verizon`



(a) Reference to Patreon during a Video



(b) Reference to Patreon in the Video Description

Figure 3.1: References by a YouTube Content Creator (Kurzgesagt) to their Patreon Webpage

creators who simultaneously use both YouTube and Patreon and those who operate only on Patreon (not YouTube).

To inform our empirical strategy, we start by developing a simple theoretical model that studies how CCs allocate their efforts across two potential platforms that offer different monetization features: YouTube, with an ad-based monetization, and Patreon, which relies on users' monthly subscriptions. On both platforms, CCs cater to audiences with different preferences for content toxicity. At the same time, YouTube's content moderation policy can change. It can be lenient, where content toxicity has no direct impact on a creator's revenue, or it can be more controlling, where a creator's content toxicity can negatively impact their revenue. In contrast, the level of toxicity on Patreon does not have a direct impact on revenue since contents are funded and displayed privately. Our theoretical framework shows that in response to YouTube's shift from a lenient to a more controlling moderation policy, CCs redirect their efforts and engagement to Patreon. As a direct result, the level of toxicity decreases on YouTube and increases on Patreon.

Following the same approach as the model, we examine Patreon CCs in the "Video" category from August 2017 to August 2018. We focus on CCs who multi-home (participate on both Youtube and Patreon) and thus are affected by YouTube's content moderation. Using a difference-in-differences (DiD) design, we estimate the causal effect of YouTube's content moderation shock by comparing these multi-homing creators to those who are never active on YouTube (and thus form a control group for our analysis). To do this, we use data from Graphtreon, a site that regularly scrapes Patreon, and we supplement this data with our own scraping of Patreon.

With this empirical design, we validate our model's predictions and uncover additional insights relevant to platforms, creators, and policymakers. First, we find that in the aftermath of the YouTube "Adpocalypse", CCs who engage in multi-homing increase the amount of subscriber-only content they offer on Patreon. In addition, the number of Patreon subscriptions increases. These findings are consistent with the notion that following YouTube's moderation policy change, CCs are motivated to shift their efforts and user base to Patreon as its relative profitability increases. This observation is consistent with the study by El-Komboz, Kerkhof, and Loh 2023, which examines how the same shock affects YouTube CCs. In addition, our research shows that the number of likes on subscriber-only content also increases, suggesting that multi-homing CCs not only produce a greater quantity of content, but also improve its quality.

For platforms, these results suggest that the different monetization strategies for CCs present a form a platform differentiation (similar to a result found in Casner and Teh 2022). For an incumbent platform, implementing restrictions to the existing monetization strategy will damage the content creator side of the market and this can reduce consumer engagement through the indirect network effects that connect the two groups. However, these losses could be mitigated by simultaneously implementing alternative monetization strategies; for example, if YouTube had also implemented a subscription model at the time of the policy change that does not follow the same restrictions required to advertise.[42] For CCs, our results highlight how multi-homing increases flexibility in monetization which can insulate profits from negative platform level shocks.

We also measure content toxicity by analyzing the titles of the Patreon content employing pre-trained algorithms provided by Jigsaw and Google's Counter Abuse Technology team (Mondal, Silva, and Benevenuto 2017; ElSherief et al. 2018; Han and Tsvetkov 2020). Doing so, we can measure toxicity of free contents (available also on other platforms such as YouTube) and subscriber-only contents and this helps us understand how the toxicity of different content evolves before and after the YouTube "Adpocalypse". Our results suggest that the overall toxicity of all subscriber-only contents (only accessible via Patreon) increases after the shock. Accordingly, and confirming our theory, toxicity migrates to Patreon in response to YouTube's content restrictions. This offers entrant platforms another avenue to increase growth as smaller platforms are less likely to come under scrutiny for such content (until they become sufficiently large). From a policy perspective, this suggests that targeting large online companies may not actually eliminate internet wide toxicity as the creators and viewers can migrate to another platform. In fact, this migration of controversial content to another platform partially mirrors the emergence of the platform "Truth Social" following the Donald Trump Twitter ban.

3.2 Literature Review and Contribution

Broadly speaking, our work contributes to the literature on platform competition in two-sided markets, where CCs generate income through consumer involvement and thus create the indirect network effects that connect the two groups. However, this literature, founded by Rochet and

[42] YouTube has a "membership program" that is similar to a subscription service, but its restrictions are consistent with the requirements for earning revenue from advertising.

Tirole 2003, Parker and Van Alstyne 2005, and Armstrong 2006, often assumes that platforms are symmetric in size and monetization strategies which is not the case in our context. Furthermore, this literature often concludes that multi-homing users are worse-off than single-homing users, in contrast with our findings.

Only a few empirical studies examine cases of competition between asymmetric platforms. For example, Farronato and Fradkin 2022 reveal that the differentiation between Airbnb listings and hotel rooms can result in substantial welfare gains from home-sharing when cities host large events that generate an influx of tourism beyond the local hotel capacity. This highlights how a platform entrant that differentiates itself from the industry incumbent is better positioned to increase its market share. Li and Zhu 2021 consider how daily deal platforms might induce high performance users on a rival platform to multi-home and how platforms try to prevent other platforms from inducing their users to multi-home. While no such actions were taken by YouTube or Patreon in our setting, our work complements theirs by revealing how alternative business models can also play an important role in benefiting the users that multi-home.

One paper that closely relates to ours is El-Komboz, Kerkhof, and Loh 2023 who consider the direct impact of the YouTube "Adpocalypse'" on YouTube CCs. Using a regression discontinuity design, they found that CCs who no longer qualify for platform monetization channels produce both fewer pieces of content and lower quality content compared to creators who remained eligible. Although they do not consider platform competition or creator multi-homing, their results are in line with our own findings, indicating a shift in content creator efforts from YouTube to Patreon in response to YouTube's monetization restrictions. Another paper that is similar to ours is Kesler 2022. He shows that Apple's change in iOS policy which reduced third-party apps' ability to track consumer locations results in third-party apps pivoting their monetization strategies away from within app advertising towards payment for or within apps. Combined with our findings, these results suggest that flexibility in monetization strategies by third-party contributors can help improve third-party profitability in the face of policy restrictions by the platform.

Naturally, our work on CCs contributes to the literature on platforms that manage user-generated content.[43] For example, Sen et al. 2023 determine how different moderation strategies impact platform content when content production occurs through professionals and through

[43] See Luca 2015 for an overview of the literature on user-generated content.

user-generated content. Similarly, Paridar, Ameri, and Honka 2023 consider how monetary and non-monetary rewards can incentivize specific types of content on a user-generated content platform for board games. Unlike the two previous studies (and similar to our setting), Wlömert et al. 2024 take an across platform approach to consider how previously protected music that becomes available on a platform that hosts user-generated content can reduce artist and music producer revenues through demand cannibalization. We contribute to this literature by revealing how a content moderation policy on one platform will impact the production and consumption of user-generated content on another platform that offers CCs with an alternative way to monetize their content.

Lastly, our paper contributes to the strand of literature analyzing content moderation and the development of harmful online content. Content moderation has been shown to be effective in reducing harmful content, being through voluntary platform measures (Chandrasekharan et al. 2017, Srinivasan et al. 2019), as well as stipulated by law (Andres and Slivko 2021a). However, most studies stay on the same platform when analyzing content moderation efforts. Remarkable examples include Ali et al. 2021 and Rauchfleisch and Kaiser 2021, who analyze the phenomenon of deplatforming, i.e. users migrating to another platform with laxer moderation efforts. Both studies find that toxic users tend to migrate or multi-home to other platforms, and even increase their toxicity on these platforms. Alternatively, Bhargava 2023 considers how taxing advertising-based platforms to subsidize subscription-based platforms impacts platform content moderation practices. In line with these studies, our paper documents an increase in toxicity on the platform Patreon triggered by content moderation efforts on YouTube.[44]

3.3 Theoretical Framework

Consider a simple model for how CCs allocate their content efforts across two potential platforms (YouTube and Patreon) that have different monetization features. We assume that the cost of content production is zero but that every CC is limited in the amount of effort they can exert to develop content across platforms. More specifically, let E denote the total amount of effort that a

[44] Lefouili and Madio 2022 and Teh 2022 consider the platform's perspective of moderating the creator side and how such restrictions might impact the overall quality of the platform, creator competition within the platform, and platform liability from content that harms society.

CC can offer, let e_{YT} denote the amount of effort toward YouTube, and let e_P denote the amount of effort toward Patreon so that $E = e_{YT} + e_P$.[45]

In terms of revenues, YouTube and Patreon have different business models and pay their CCs differently. YouTube pays for content through advertising revenues that are split between YouTube and its CCs. Let $r_{YT}(t_{YT}, \mathbb{B})$ capture the marginal advertising revenue from one unit of effort on YouTube, where t_{YT} denotes the level of toxicity of the CC on YouTube and $\mathbb{B} \in \{0,1\}$ denotes the leniency of the YouTube moderation policy. If $\mathbb{B} = 0$, then the YouTube moderation policy is lenient, and t_{YT} has no impact on r_{YT}. If $\mathbb{B} = 1$, YouTube is more likely to control and demonetize toxic content. Thus, a greater t_{YT} reduces r_{YT}. In practice, CCs have control over the content that they post; at the same time, CCs differ across audience toxicity preferences. To model these features, suppose that each CC receives a toxicity draw given by $\tau \sim F(\cdot)$ so that deviations away from this toxicity level are costly to the CC in the form of $(\tau - t_{YT})^2$ for content on YouTube.

On Patreon, CCs offer a menu of services for different membership fees that are paid directly by consumers. The ability to offer a menu of services allows CCs to price discriminate so that the first few units of effort on Patreon go toward the most profitable forms of content. To model this environment simply, suppose that the marginal revenue from effort diminishes on Patreon so that the marginal revenue at effort e is given by $r_P(e) = \rho - e$, where $\rho \sim G(\cdot)$ is a random draw that captures the maximum revenue generated from initial effort on Patreon which, like toxicity τ, can differ across CCs.[46] This implies that a CC's total revenue from Patreon when exerting effort level e_P is given by $\int_0^{e_P} (\rho - e) de$. Note that toxicity does not impact revenues on Patreon because Patreon is a pure "membership portal" that, to this point, has not faced scrutiny regarding toxicity on its platform. Much like on YouTube, we allow for CCs to choose the toxicity level on Patreon so that the cost of setting toxicity t_P on Patreon is given by $(\tau - t_P)^2$ for content on Patreon.[47]

[45] Note that the main predictions hold if we allow CCs to differ in the total amount of effort they exert.

[46] Naturally, CCs are also heterogenous in the amount of views they earn on YouTube. We avoid heterogeneity on YouTube so that we can make comparative statics with respect to r_{YT}. However, additionally allowing for YouTube heterogeneity does not change the predictions of our model.

[47] Alternatively, each content could vary in its toxicity draw or costs from altering the toxicity draw could depend on substitution in toxicity across platforms. Incorporating these features into the model does not qualitatively change our main results.

Combining the revenue streams from YouTube and Patreon, we see that the CC maximization problem is given by

$$\begin{aligned} \max_{e_{YT},e_P,t_{YT},t_P} \quad & r_{YT}(t_{YT},\mathbb{B})\cdot e_{YT} - (\tau - t_{YT})^2 + \int_0^{e_P} (\rho - e)de - (\tau - t_P)^2 \\ & \text{s.t. } E = e_{YT} + e_P. \end{aligned}$$

Solving allows us to determine equilibrium effort levels across platforms as well as CC level toxicity across platforms. In terms of effort, we have the following result:

Proposition 1. *In equilibrium, CC effort levels are given by* $e_P^* = \max\{0, \min\{E, \rho - r_{YT}\}\}$ *and* $e_{YT}^* = \max\{0, \min\{E, E + r_{YT} - \rho\}\}$.

In other words, an interior solution occurs with $e_P^* = \rho - r_{YT}$ whenever $(\rho - r_{YT}) \in (0, E)$. In this case, we have that the equilibrium marginal revenue from effort toward Patreon, given by $\rho - e_P^*$, equals the marginal revenue from YouTube effort, r_{YT}, so that the CC produces content on both platforms. Instead, if the Patreon marginal revenues are too low ($\rho < r_{YT}$), then the CC only produces on YouTube; similarly, if the Patreon marginal revenues are too high ($\rho - E > r_{YT}$), then the CC only produces on Patreon.

Our model provides some natural predictions for how different types of CCs might respond to the YouTube ad boycott. To highlight these predictions in a manner that aligns with our empirical approach, we compare the equilibrium in the pre- and post-boycott periods. Regarding effort responses to the YouTube ad boycott, we see that efforts migrate from YouTube to Patreon:

Corollary 1. *Content creator efforts migrate from YouTube toward Patreon following a YouTube ad boycott:* $\frac{de_P^*}{dr_{YT}} \in \{0, -1\}$ *and* $\frac{de_{YT}^*}{dr_{YT}} \in \{0, 1\}$.

To determine how toxicity changes across platforms, we must measure toxicity at the CC-platform level. In other words, toxicity on YouTube and Patreon are given by

$$\begin{aligned} T_{YT} &= e_{YT} \cdot t_{YT}, \\ T_P &= e_P \cdot t_P. \end{aligned}$$

Thus, in terms of toxicity at the CC-platform level, we see that

Proposition 2. *In equilibrium,* $T_{YT}(\mathbb{B}=0) \geq T_{YT}(\mathbb{B}=1)$ *and* $T_P(0) \leq T_P(1)$.

Hence, when toxic behavior becomes more costly on YouTube, toxicity migrates from YouTube to Patreon.

3.4 The Empirical Setting

Our empirical analysis follows a methodology similar to the approach outlined in the model of Section 3.3. Specifically, we direct our focus towards two distinct groups of Patreon CCs: those who multi-home on Patreon and YouTube and those who operate on Patreon without multi-homing on YouTube. To accomplish this, we use a dataset sourced from Graphtreon, a data provider that routinely collects information from Patreon. This dataset provides us with a comprehensive overview of the number of CCs on the platform over time. It includes data on whether they engage in multi-homing on YouTube and the total count of the paying subscribers for each creator active on the platform over time. However, this dataset does not include specific details about the content produced by each creator. To address this limitation, we enrich this dataset by conducting our own data scraping directly from Patreon to gather content-related information.

In the following part of this section, we start presenting contextual information relevant to our analysis, including details about YouTube and Patreon, the YouTube Partner Program, and the chronological sequence of events related to the so-called "Adpocalypse" shock. Following this contextual overview, we provide more comprehensive descriptive statistics derived from the dataset employed in our analysis.

3.4.1 YouTube and Patreon

YouTube is one of the world's leading video-sharing platforms, with a vast array of user-generated content spanning a wide range of genres, from entertainment and education to gaming and vlogging. Founded in 2005, YouTube has since become one of the major platforms in the digital world. The platform enables users to upload, view, and share videos. YouTube currently boasts over 2 billion logged-in monthly users, with millions of videos uploaded daily.[48]

[48] For more information about updated YouTube statistics, see Statista: `https://www.statista.com/topics/2019/YouTube`

YouTube offers a range of monetization options for CCs, allowing them to earn revenue from their videos. The most common method of monetization on YouTube is through advertising revenue. CCs can enable ads on their videos, and they earn a share of the revenue generated from ads shown to viewers. Moreover, CCs can offer channel memberships to their audience. Viewers who become channel members pay a monthly fee and, in return, receive perks like custom badges, emojis, and exclusive content.

To unlock the various monetization features offered by the platform, CCs must officially partner with YouTube and enroll in the YouTube Partner Program (YTPP). In the following subsection, we will explore the specific criteria that creators need to satisfy to qualify for YTPP, and how these requirements have evolved over time. Apart from the monetization features designed for CCs on the YouTube platform, creators can also sponsor content or use crowdfunding platforms like Patreon or external donation links to receive direct financial support from their fans.

Patreon is a platform that enables CCs to receive direct financial support from their dedicated fans and followers. It empowers creators to establish a direct connection with their audience by offering exclusive, often more private, content in exchange for monthly subscriptions. That is, unlike a Youtube "subscriber" who does not necessarily pay anything to Youtube or the CC, a "patron" on Patreon is paying a monthly subscription for which the majority of the at least $1 membership goes directly to the CC.[49] Patreon is one of the most renowned crowdfunding platforms, especially within the realm of online content creation, including YouTubers, podcasters, and other digital creators. In the context of YouTube, CCs can reference their Patreon webpage directly in their videos or within the video description, as illustrated in Figure 3.1. This approach allows CCs to diversify their income streams and reduce their reliance on YouTube's advertising revenue, mitigating the impact of changes in YouTube's policies, such as ad monetization eligibility.

3.4.2 The YouTube "Adpocalypse"

In early 2017, major advertisers and prominent brands, including well-known companies such as Coca Cola, Amazon, and Walmart, became concerned about their advertisements appearing

[49] Patreon only takes between 5-12% of the monthly subscription revenues generated from the CC's patrons.

alongside content they deemed inappropriate or extremist.[50] In response, many of these brands temporarily suspended their advertising on YouTube. This advertising boycott, and more generally, this period in YouTube's history, is known as the YouTube "Adpocalypse".

Before the "Adpocalypse" of 2017, YouTube's approach to content moderation was considerably more relaxed compared to what it would become in the aftermath of that event. YouTube primarily relied on algorithms to monitor and moderate content. These algorithms were designed to flag and remove content that violated YouTube's Community Guidelines, including policies against hate speech, violence, explicit content, and copyright infringement. However, the accuracy of these algorithms was not flawless, and inappropriate or borderline content often evaded detection.[51] In addition to algorithmic moderation, YouTube heavily depended on its user reporting system. Users were encouraged to report content they found offensive or in violation of guidelines. This crowd-sourced approach to moderation meant that content could be flagged based on user reports, and YouTube would then review the reported content. While this system helped identify problematic content, it also resulted in inconsistencies as flagged content was reviewed by human moderators and subject to their interpretation.

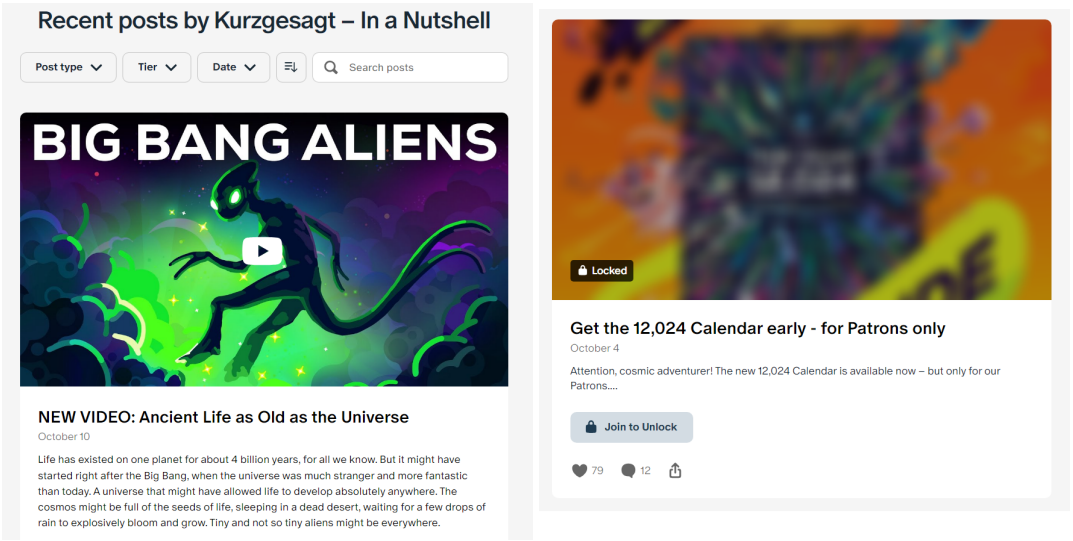
Before the "Adpocalypse", YouTube's monetization model was relatively lenient. CCs could monetize their videos with ads, and there were fewer restrictions on what content could be monetized. In response to the events in 2017, YouTube implemented a series of demonetization policies aimed at addressing concerns raised by advertisers and improving brand safety on the platform. Quoting the former YouTube CEO from an online post in December 2017:[52] "We want advertisers to have peace of mind that their ads are running alongside content that reflects their brand's values. Equally, we want to give creators confidence that their revenue won't be hurt by the actions of bad actors."

YouTube introduced more comprehensive and explicit content guidelines to provide CCs with a better understanding of what was considered acceptable on the platform. Moreover, YouTube

[50] For specific cases, refer to these articles: `https://www.thetimes.co.uk/article/big-brands-fund-terror-knnxfgb98`, `https://www.wsj.com/articles/disney-severs-ties-with-YouTube-star-pewdiepie-after-anti-semitic-posts-1487034533`. For a general overview, see The Verge: `https://www.theverge.com/2019/4/5/18287318/YouTube-logan-paul-pewdiepie-demonetization-adpocalypse-premium-influencers-creators`

[51] Polygon Article: `https://www.polygon.com/2017/11/7/16620400/YouTube-algorithm-kids-programming`

[52] For the entire post, see YouTube blog: `https://blog.YouTube/news-and-events/expanding-our-work-against-abuse-of-our/`



(a) Public Post (b) Patrons-Only Post

Figure 3.2: Observable Posts on Patreon (Kurzgesagt)

modified the YouTube Partner Program (YTPP) on February 20th, 2018, which previously had almost no restrictions on CCs. Now, "(YouTube) channels with fewer than 1,000 subscribers or 4,000 watch hours will no longer be able to earn money on YouTube".[53] Moreover, channels had to comply with YouTube's Community Guidelines and Ad Policies. This meant that the content on the channel had to adhere to YouTube's rules regarding hate speech, violence, explicit content, and other policy areas. Channels with repeated violations could face demonetization or other penalties.

These changes in the YTPP not only set explicit eligibility standards but also revealed a nuanced shift in YouTube's approach to content moderation. They implied that CCs with smaller audiences might be too numerous to moderate effectively and, therefore, would not be monetized. On the other hand, creators with larger followings would have the opportunity for monetization but would need to navigate the scrutiny of the platform's content moderation policies.

3.4.3 Graphtreon and Patreon Data

We trace CCs' content on Patreon with the help of data from Graphtreon, a website that tracks the overall number and characteristics over time of CCs on Patreon with at least one patron (subscriber). In our analysis, the dataset is composed of Patreon snapshots from August 2017 to August 2018 for the "Video" category. We combine all these snapshots to create an unbalanced

[53] `https://blog.YouTube/news-and-events/additional-changes-to-YouTube-partner/`

panel dataset. In each snapshot, we observe CCs if they are present on the Patreon website on the snapshot date and have at least one patron. Consequently, for each CC in the dataset, we identify entry and exit events, as well as several CC characteristics. Some of these characteristics remain constant over time, such as the CCs' category and the Patreon web link, while others are updated with each snapshot. These dynamic characteristics include the number of patrons subscribed to each CC, monthly earnings, and whether the CC has a direct web link to other platforms like YouTube.

We enhance the Graphtreon dataset by extracting information directly from Patreon. Specifically, we leverage the direct links to Patreon provided by each CC. Not all the Patreon CCs listed in the Graphtreon dataset were still active on the platform when we began our scraping in 2022. Therefore, our analysis is limited to those CCs that were active throughout the year 2017 and whose webpages had not been deleted as of Spring 2022.

Since Patreon is a subscription-only platform, we do not have direct access to all the content posted by each CC, such as their videos (see Figure 3.2). However, we are able to observe important details, including the posting date, title, the number of comments and likes for each post, and whether the content is freely accessible or requires a subscription. Consequently, for each month in our tracking period, we collect data on the number of free or subscribers-only content items produced by CCs, as well as the number of likes and comments received by each piece of content.

Furthermore, through analysis of the content titles, we employ pre-trained algorithms provided by Jigsaw and Google's Counter Abuse Technology team (Mondal, Silva, and Benevenuto 2017; ElSherief et al. 2018; Han and Tsvetkov 2020) to identify and measure the level of toxicity in the content produced by each CC over time. Before delving into the details of our identification design, we provide an overview of the dataset we have gathered regarding Patreon CCs (CCs). In Tables 3.1 and 3.2, we present information regarding the characteristics of CCs on Patreon. Our dataset is limited to the period from August 2017 to August 2018, focusing on CCs for whom we collected information when we scraped their Patreon webpages in Spring 2022.

Table 3.1 displays key statistics for the CCs within our sample. We have data for over 11,000 CCs, with an average patron count exceeding 80. Not all CCs disclose their earnings and our sample reveals that roughly 80% of the CCs choose to reveal. For those who do disclose their

earnings, the average monthly income is more than 320 US$, which suggests that, on average, each patron contributes around 4 US$ per month in revenue. Many CCs on Patreon also maintain a presence on YouTube. Specifically, 85% of CCs have, at least once during the period of interest, provided a link to YouTube. Regarding the volume of content, comments, and likes they receive each month, CCs produce an average of four pieces of content per month. These pieces of content receive an average of 18 comments and 41 likes. This indicates that CCs on Patreon enjoy an actively engaged audience of patrons who interact with their content.

Table 3.1: Patreon CC Characterstics

	N	Mean	Median	SD	Min	Max
Patrons	11371	84	10	369	1	12,206
Earnings	9465	321	54	1,176	.0092	38,878
YT Presence	11371	.85	1	.36	0	1
# Contents/Month	11371	4.1	1.5	7.5	0	170
# Comments/Month	11371	18	.77	80	0	2,076
# Likes/Month	11371	41	1.8	190	0	6,155

Notes. The dataset is sourced from Graphtreon and our independent web scraping of Patreon webpages. We narrow our focus to CCs who are active on Patreon (category "Video") for at least one month within the period spanning from August 2017 to August 2018.

In Table 3.2, we present the same dataset, but this time we no longer average the data at the individual CC level. Instead, we consider each individual CC-month observation, which is the dimension we use in our identification design. As previously mentioned, our dataset comprises more than 11,000 unique CCs and a total of 131,000 CC-month observations. To reiterate, on average, CCs produce four pieces of content per month. These content pieces can be categorized into three types: free content, base content, which requires a minimum subscription (one dollar) to access, and premium content, which requires a subscription of more than one dollar for patrons to view.

In the upcoming sections, we will merge base and premium content and solely focus on analyzing the distinctions between free and paid content. However, it is worth noting that these three types of content (free, base, and premium) constitute a similar proportion of the total number of content pieces. Content posted on Patreon typically includes images and videos, as evidenced by the fact that, on average, CCs post more than 3 images and nearly two videos every month.

Table 3.2: Patreon CC-Month Observations: Types of Contents

	N	Mean	Median	SD	Min	Max
# Contents	131006	4.01	1	8.24	0	271
# Free Contents	131006	1.34	0	3.90	0	162
# Base Contents	131006	1.55	0	4.32	0	201
# Premium Contents	131006	1.11	0	4.18	0	252
# Images	131006	3.18	0	7.23	0	230
# Videos	131006	1.89	0	5.13	0	216

Notes. The dataset is sourced from Graphtreon and our independent web scraping of Patreon webpages. We narrow our focus to CCs who are active on Patreon (category "Video") for at least one month within the period spanning from August 2017 to August 2018.

Before we delve into our identification strategy, we provide an overview of how the number of CCs evolved around the time when YouTube implemented a stricter content moderation policy. In Figure 3.3, we present a graphical representation of the number of CCs in the "Video" category that either joined or left Patreon between August 2017 and August 2018. As can be seen in the figure, the platform experienced growth during this period, with more new CCs joining than leaving. Notably, we can see specific spikes in the months of August 2017, November 2017, and January 2018, which coincide with the timing of the YouTube ad boycott. Additionally, throughout 2018, the influx of new CCs has exceeded the levels seen in the last few months of 2017. However, despite these fluctuations in new entrants, the total number of CCs on Patreon remained relatively stable. The total number of CCs present from August 2017 to August 2018 exceeded 11,000, while the number of new CCs joining the platforms in late 2017 or early 2018 is less than 1200. This suggests that while there were fluctuations in the number of new arrivals, the overall CC population on Patreon experienced only modest changes.

3.5 Identification Strategy

With our empirical design, we aim to study how CCs respond to the changes in the YouTube monetization policy due to the "Adpocalypse" ad boycott. To do so, we investigate the evolution in the quantity and nature of content produced, as well as changes in the number of patrons that pay monthly subscriptions to two sets of Patreon CCs from August 2017 to August 2018. We focus on CCs that were multi-homing Patreon and YouTube, as they constitute the "treated" group directly affected by the introduction of the YouTube moderation policy. We identify these

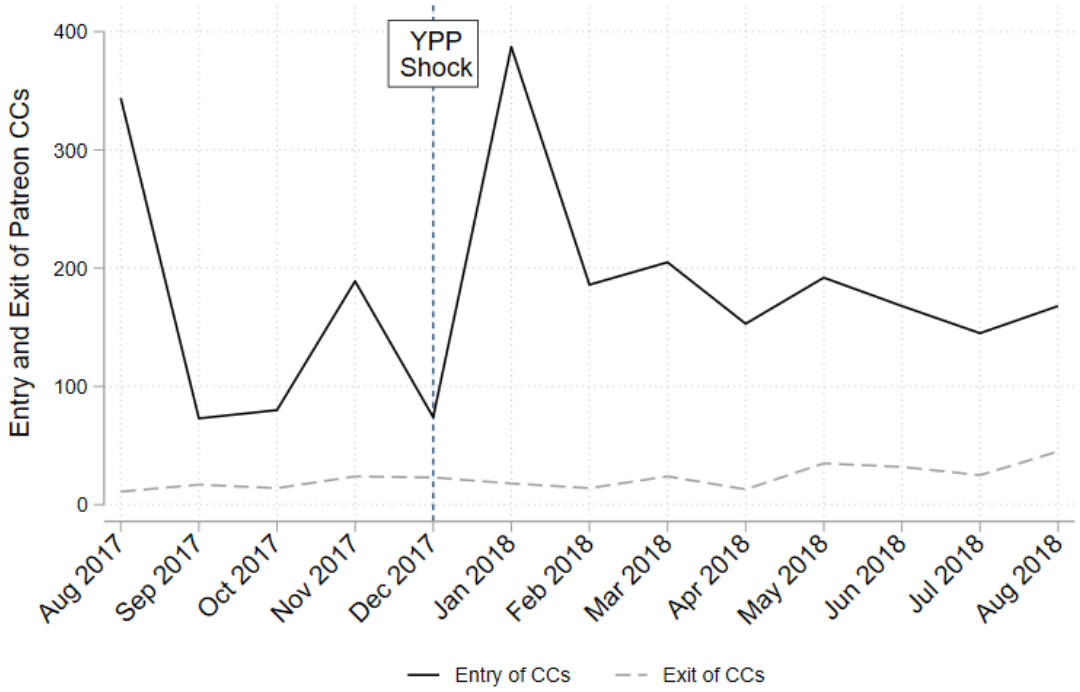


Figure 3.3: Entry and Exit on Patreon Before and After the YouTube "Adpocalypse"

Notes: The dataset is sourced from Graphtreon and our independent web scraping of Patreon webpages. We narrow our focus to CCs who are active on Patreon (category "Video") for at least one month within the period spanning from August 2017 to August 2018.

CCs based on whether they display a YouTube link on their Patreon webpage at least once during the period of analysis. Conversely, Patreon CCs without YouTube links on their webpages serve as our control group, as they remain unaffected by the YouTube policy changes.

The change in YouTube's moderation policy, specifically the tightened eligibility criteria for the YouTube Partner Program, was formally implemented in February of 2018.[54] However, the shift in YouTube's moderation approach began several months earlier during the "Adpocalypse" of 2017 and the resulting ad boycotts. We cannot trace back the exact moment when YouTube altered its approach and CCs became aware of this change. However, in December 2017, the former YouTube CEO, Susan Wojcicki, officially announced expanded efforts to combat abuse on the platform.[55] Thus, we use December 2017 as the month when CCs became aware of the forthcoming changes. Some CCs may have anticipated the changes prior to this date, but we consider that December 2017 represents the time when most CCs started taking measures to respond to the upcoming moderation policy.

[54] `https://blog.YouTube/news-and-events/additional-changes-to-YouTube-partner/`
[55] `https://blog.YouTube/news-and-events/expanding-our-work-against-abuse-of-our/`

Thus, we use a Difference-in-Difference (DiD) identification approach and compare Patreon CCs in the "treated" group (multi-homing on YouTube) and "control" group (not multi-homing on YouTube) from August 2017 to August 2018 before and after December 2017. The estimating regression to capture the causal impact of the moderation policy on Patreon CCs' activity is:

$$y_{it} = \alpha_i + \rho_t + \beta_1 YT_i + \beta_2 After_t + \beta_3 YT_i \times After_t + \varepsilon_{it}, \tag{3.5.1}$$

where y_{it} denotes the variable of interest across different dimensions of CCs' activity affected by the YouTube policy (e.g., number of subscribers or contents), α_i and ρ_t are the CC and time fixed effects, YT_i is a dummy variable equal to 1 if the CC i has multi-homed on Patreon and YouTube for at least one month during the period of analysis and equal to 0 if the CC i has never multi-homed on YouTube, and $After_t$ is equal to 1 for all snapshots after December 2017 and equal to 0 otherwise. The coefficient β_3 captures the impact of the change in the YouTube monetization policy under the assumption that Patreon CCs that do not multi-home on YouTube provide a good counterfactual group for the evolution of the variables of interest of the CCs that are active on Patreon and YouTube.

In the next section, we present the results of the Difference-in-Differences (DiD) approach for all the variables of interest. Additionally, for each variable, we check for the absence of potential pre-trends between the different groups of CCs using an event-study approach. We use the following lead-lag model in which y_{it} is regressed on the product of the dummy variable YT_i and a comprehensive set of dummy variables for each snapshot. The model controls for CC fixed effects and time fixed effects:

$$y_{it} = \alpha_i + \rho_t + \sum_{\tau=Sep17}^{Jul18} \beta_\tau YT_i \times 1(t=\tau) + \varepsilon_{it}. \tag{3.5.2}$$

In all graphs, the coefficients corresponding to months before December 2017 are approximately zero and do not display a discernible trend. Consequently, the evolution of the variables of interest for the treated and control CCs appeared to be similar before the moderation policy change. This observation supports the parallel trend assumption and confirms the validity of our empirical design.

3.6 Results

Using the DiD identification design outlined in Section 3.5, we can study different dimensions related to the number of patron subscriptions and the quantity and the nature of the contents of Patreon CCs. Before focusing on each variable, here we provide an overview of the main results.

CCs multi-homing on YouTube produce more content that is exclusively available to their subscribers on Patreon. Instead, the quantity of the public content remains unchanged. Together, these results suggest that CCs affected by the moderation policy change are prioritizing the now more attractive monetization strategy offered by Patreon. Also the number of patrons subscribing to Patreon increases, possibly due to active recommendations from CCs encouraging viewers to transition from YouTube to Patreon. This rise in the number of subscribers also has a significant impact on earnings.

Moreover, we can show that treated CCs are also investing more effort on Patreon, not only in terms of quantity (more content) but also in the quality of content. Notably, we observe an increase in the number of likes on Patreon for those CCs affected by the content moderation policy compared to CCs who do not multi-home on YouTube. The increase in the number of likes is primarily noticeable for subscription-only content, and this positive effect remains robust even after controlling for changes in the number of content and patrons. It is essential to emphasize this point, as one might anticipate that more content and more patrons would mechanically increase the number of likes. A similar effect is evident for comments, with more comments on subscribers-only content posted on Patreon. However, once we account for the number of content pieces and the number of patrons, the effect becomes null, suggesting that changes in the number of comments are primarily linked to increases in subscribers and content volume. Finally, the overall toxicity of Patreon increases, mainly due to CCs posting more content on this platform. However, there is no observed increase in the average toxicity of the content itself.

In the remainder of this section, our focus will shift to each individual result, where we will describe the DiD table and the event study figures in detail.

3.6.1 Number of Patrons and Earnings

We first investigate whether the measurable outcomes of CCs who engage in multi-homing on YouTube, change on Patreon compared to those CCs who do not multi-home. Table 3.3

documents a significant and strong increase in the number of patrons subscribing to multi-homing CCs. In line with these findings, CCs' earnings on Patreon experience a notable surge following the implementation of the content moderation, driven by the substantial increase in the number of patron subscriptions. These results show that CCs are succesful in making up for lost or threatened YouTube profits on the platform Patreon.

To visualize these effects, Figure 3.4 in Appendix C presents the event study for the logarithm of the number of patrons and earnings. In both cases, a clear and positive effect becomes evident in the post-YTPP period. However, it is worth noting that we may also observe some degree of anticipation among CCs, as they appear to have started attracting users to Patreon even before December 2017, possibly in anticipation of the impending tightening of rules.

Table 3.3: Difference-in-Differences: Log of Number of Patrons and Earnings

	Log Patrons	Log Earnings
$post_t^{Dec2017} \times YT_i$	0.12***	0.13***
	(0.01)	(0.02)
User FE	✓	✓
Time FE	✓	✓
Mean	2.72	4.03
R^2	.968	.95
N	108,315	86,057

Notes: Standard errors clustered by CC are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.6.2 Number of Contents

Next, we zoom in on the question of *how* affected CCs achieve to increase their performance on Patreon. We first investigate the impact of the YTPP on the content generated by CCs on Patreon. In this context, the outcome variable, denoted as y_{it} in Equations 3.5.1 and 3.5.2, is defined either as a binary variable indicating whether a CC produced at least one piece of content in a given month (as shown Columns 1-3 in Table 3.4) or as the logarithm of the number of content items generated by each CC within that month (Columns 4-6 in Table 3.4). These two settings correspond to the extensive and intensive margins of the effect. While the first setting looks at whether content moderation leads to CCs posting more frequently in more months, the second establishes whether CCs create a higher number of contents per month. In both settings,

Table 3.4: Difference-in-Differences: Presence and Number of Contents

	Presence of Content			Log No. of Contents		
	All	Free	Paid	All	Free	Paid
$post_t^{Dec2017} \times YT_i$	0.00	0.01*	0.01	0.08***	-0.02	0.07***
	(0.01)	(0.01)	(0.01)	(0.02)	(0.03)	(0.03)
User FE	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓
Mean	.533	.32	.403	1.39	.929	1.24
R^2	0.61	0.53	0.64	0.72	0.66	0.70
N	108315	108315	108315	57006	33394	42886

Notes: Standard errors clustered by CC are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

our analysis considers all CCs active on Patreon in the "Video" category, and we focus on those who joined the platform before the implementation of the YTPP and remained active afterwards.

The findings from Table 3.4 reveal a significant trend: CCs who engage in multi-homing on YouTube post a greater volume of content compared to their non-multi-homing counterparts on the intensive: They do not generate content in more months, but they do generate a higher number of contents within a month. Moreover, this effect is primarily driven by the production of paid content, which is accessible exclusively through subscription, as opposed to free content that is available to all users.

To visualize these trends, Figure 3.5 in Appendix C showcases three event study graphs utilizing the logarithm of the number of total content items, free content, and paid content. All three graphs demonstrate a lack of noticeable pre-trend behavior before the YTPP implementation. However, an effect emerges after January 2018 concerning the production of content and paid content.

In order to tackle potential concerns about the specific timing of when the impact of YouTube's moderation changes should come through on Patreon, we rerun the above analyses using different post cutoffs. Tables 3.8 and 3.9 in Appendix B confirm the previous findings: Multi-homing CCs produce more content in the intensive margin, and this increase is primarily driven by an increased number of paid contents.

3.6.3 Number of Likes

Following our examination of the cross-platform impact on the quantity of content and the total number of subscribed users, we now delve into its effect on content quality, specifically measured by the number of likes received by CCs and month. The first three columns in Table 3.5 replicate the analysis previously presented in Table 3.4, focusing on the impact of the rule change on the logarithm of the number of likes for all content, free content, and paid content, employing the same specifications as outlined in Equation 3.5.1. Notably, the results demonstrate that the increase in the number of likes is indeed evident, with a noteworthy effect primarily driven by likes for paid content.

While these results are intriguing, it is reasonable to question whether these effects are merely a mechanical consequence of the increased volume of content and patrons. One might expect that more content would naturally lead to more likes, and an expanded patron base might also correlate with increased likes. To address this concern, the second half of Table 3.5 employs a classical DiD design while controlling for both the total number of patrons and the quantity of content produced in a given month. This approach departs somewhat from the classical DiD design, as we typically would not control for a variable affected by the very shock under investigation. However, we believe it is instructive to demonstrate that the rise in likes appears only partially linked to these other variables. Remarkably, Columns 4-6 of Table 3.5 reveal a significant increase in the number of likes for all content, driven by paid content, even after accounting for the mentioned control variables.

We acknowledge that the variation in likes may also depend on the characteristics of the patrons, who may vary in their inclination to leave a like. Nonetheless, these findings suggest that the YTPP's impact on likes extends beyond a simple mechanical effect and is indicative of genuine shifts in user engagement and content quality. As before, we visualize these effects in Figure 3.6 in C where we presents the event study for the logarithm of the number of likes for different types of content. In all cases, a clear and positive effect becomes evident in the post-shock period for the likes on all content and paid content.

Table 3.5: Difference-in-Differences: Log of Number of Likes

	Unconditioned			Conditioned		
	All Contents	Free Contents	Paid Contents	All Contents	Free Contents	Paid Contents
$post_t^{Dec2017} \times YT_i$	0.14***	0.05	0.15***	0.08***	0.01	0.10***
	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.03)
User FE	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓
No. Patrons				✓	✓	✓
No. Contents				✓	✓	✓
Mean	2.73	2.28	2.7	2.73	2.28	2.7
R^2	.854	.827	.843	.888	.846	.872
N	47,642	26,875	35,686	47,642	26,875	35,686

Notes: The first three columns replicate the previous analysis in a simple DiD framework. Columns 4-6 additionally control for the number of patrons and contents. Standard errors clustered by CC are in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.6.4 Number of Comments

The engagement of users with contents can also measured by considering the number of comments generated by a content or CC. Table 3.6 provides insights into this aspect, indicating an increase in the number of comments for both all content and paid content when employing the same specification as outlined in Equation 3.5.1. However, when we reevaluate this analysis while accounting for the number of content items and subscribers in the second part of Table 3.6, we do not detect a significant effect. This observation suggests that the number of comments per content and per subscriber remains relatively stable.

Table 3.6: Difference-in-Differences: Log of Number of Comments

	Unconditioned			Conditioned		
	All Contents	Free Contents	Paid Contents	All Contents	Free Contents	Paid Contents
$post_t^{Dec2017} \times YT_i$	0.11***	0.06	0.10**	0.06	0.02	0.04
	(0.04)	(0.05)	(0.05)	(0.04)	(0.05)	(0.04)
User FE	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓
No. Patrons				✓	✓	✓
No. Contents				✓	✓	✓
Mean	2.43	1.93	2.46	2.43	1.93	2.46
R^2	.778	.707	.758	.809	.725	.786
N	39,020	19,400	30,064	39,020	19,400	30,064

Notes: The first three columns replicate the previous analysis in a simple DiD framework. Columns 4-6 additionally control for the number of patrons and contents. Standard errors clustered by CC are in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.7: Difference-in-Differences: Content Toxicity

	Log Total Toxicity			Log Average Toxicity		
	All Contents	Free Contents	Paid Contents	All Contents	Free Contents	Paid Contents
$post_t^{Dec2017} \times YT_i$	0.09***	-0.01	0.08**	0.01	0.01	0.01
	(0.03)	(0.04)	(0.04)	(0.02)	(0.03)	(0.02)
User FE	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓
Mean	-4.61	-5.12	-4.81	-6.01	-6.05	-6.05
R^2	.636	.578	.622	.421	.396	.415
N	56,832	33,315	42,794	56,832	33,315	42,794

Notes: Standard errors clustered by CC are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.6.5 Content Toxicity

We conclude our series of findings by examining the toxicity of the contents generated by CCs. Utilizing our content titles and employing the pre-trained algorithms provided by Jigsaw and Google's Counter Abuse Technology team (Perspective), we measure the toxicity of each piece of content as a continous score ranging from 0 to 1. In the first part of Table 3.7, we aggregate the toxicity levels of all content produced by each CC in a given month. Conversely, the second part of Table 3.7 employs the average toxicity level of the content. These analyses serve distinct purposes: While the total toxicity enables us to assess whether Patreon becomes more toxic, in total, following the content moderation shock on YouTube; the average toxicity measures allow us to determine whether the individual contents being produced are more toxic.

Our analysis of these two toxicity measures reveals a noteworthy outcome: CCs who engage in multi-homing on YouTube do not appear to become more toxic themselves. However, as they increase their content production, especially in the realm of paid content, the overall toxicity level on the platform experiences an upsurge.

3.7 Implications

Our findings speak directly to several players in the content creator space: incumbent and entrant platforms considering content moderation and monetization policies, CCs evaluating such policies, and legislators interested in reducing toxicity. For platforms implementing content moderation policies, our work shows that such policies may leave the platform open to competition from platforms offering alternative monetization strategies. This suggests that

the losses associated with introducing content moderation could be limited if the platform also provides alternative monetization opportunities to its CCs. More generally, we reveal an important relationship between content moderation and platform monetization that potential platform entrants should be aware of: an advertising model where content is open to the public may result in future content moderation issues that could be avoided under a membership model where content is funded by consumers and is therefore able to remain private. Naturally, a platforms' choices in monetization and moderation policies impact CCs. In particular, CCs that are able to adapt their content to multiple platform business models are better able to monetize their content, emphasizing the importance of developing a variety of content. Lastly, our work suggests that a legislator interested in reducing toxicity may only be able to reduce publicly observable toxicity when targeting content moderation on a dominant platform as entrant platforms may offer an alternative option for such content to migrate.

This connection between content moderation and platform monetization appears in other industries as well. For example, Elon Musk has speculated that changing the monetization strategy of X, formerly Twitter, so that users generating content must pay an annual fee will also serve as a content moderation policy by blocking bot users.[56] Alternatively, Instagram competing with Onlyfans relates to our work on platform competition when platforms have different monetization strategies that allow for public versus private content. More specifically, in response to Onlyfans' successful use of the membership model with private content (unlike YouTube's null response to Patreon's success), Instagram launched a similar subscription model for private content to better compete with Onlyfans. These examples, combined with our own findings, highlight how important it is for platforms to account for the interplay between content moderation and monetization.

3.8 Conclusion

Our study delves into the complex dynamics of the content creator industry, with a particular focus on the aftermath of the YouTube "Adpocalypse" and the subsequent restrictions in monetization. As we examine the shifts in CCs' strategies and consumer behavior across platforms, several key findings emerge that have implications for various stakeholders in this ecosystem.

[56] Specifically, in an X post on October 17, 2023: "read for free, but $1/year to write. It's the only way to fight bots without blocking real users."

First, our theoretical model and empirical evidence confirm that CCs respond strategically to changes in moderation policies that could potentially lead to demonetization. They diversify their monetization strategies, with an increased focus on platforms like Patreon, which offers alternative revenue streams such as monthly subscriptions. This diversification not only allows CCs to adapt to changing market conditions but also results in an improvement in the quality and quantity of content available on Patreon. This finding underscores the importance of flexibility in monetization for CCs, which can insulate their profits from platform-level shocks and ultimately enhance their ability to thrive in a rapidly evolving industry. For platforms, our research highlights the significance of offering diverse monetization options for CCs. Differentiation in monetization strategies can attract and retain a diverse range of CCs, fostering a more robust ecosystem.

Regarding the issue of toxic content, our study reveals a migration of toxicity to Patreon following the YouTube restrictions. This migration pattern suggests that efforts to target large online companies alone may not effectively reduce the internet-wide prevalence of toxicity, as creators and viewers can simply shift to other platforms with fewer content moderation measures. This phenomenon underscores the challenges faced by policymakers in addressing online toxicity and highlights the need for a more comprehensive, industry-wide approach to tackling this issue.

Our research provides valuable insights into the ever-evolving content creator industry, shedding light on how CCs, platforms, and policymakers respond to significant disruptions and challenges. As this industry continues to evolve, it will be essential for all stakeholders to adapt, innovate, and collaborate to create a safer and dynamic online content ecosystem.

Appendices

A Appendix of Proofs

Proof of Proposition 1: Considering equilibrium effort levels reduces the CC's four variable constrained maximization problem to a simple two variable Lagrange maximization problem:

$$\begin{aligned} \max_{e_{YT},e_P} \quad & r_{YT}(t_{YT},\mathbb{B})\cdot e_{YT} + \int_0^{e_P}(\rho - e)de, \\ & \text{s.t. } E = e_{YT} + e_P. \end{aligned}$$

In this case, an interior solution occurs with $e_P^* = \rho - r_{YT}$ and $e_{YT}^* = E - \rho - r_{YT}$ whenever $(\rho - r_{YT}) \in (0,E)$. Instead, if $\rho < r_{YT}$, then $e_P^* = 0$ and $e_{YT}^* = E$, and if $\rho - E > r_{YT}$), then $e_P^* = E$ and $e_{YT}^* = 0$. □

Proof of Proposition 2: If no ad boycott exists ($\mathbb{B} = 0$), then $r_{YT}(t_{YT},\mathbb{B})$ is not effected by t_{YT} so that CCs implement their default level of toxicity: $t_{YT}^* = t_P^* = \tau$. Instead, if an ad boycott exists so that $r_{YT}(t_{YT},\mathbb{B})$ is decreasing in t_{YT}, then t_{YT}^* is given by

$$0 = \frac{dr_{YT}(t_{YT},1)}{dt_{YT}}\cdot e_{YT} + 2(\tau - t_{YT}),$$

where $\frac{dr_{YT}(t_{YT},1)}{dt_{YT}} < 0$. Solving implies that

$$t_{YT}^* = \frac{dr_{YT}(t_{YT},1)}{dt_{YT}}\cdot\frac{e_{YT}}{2} + \tau < \tau.$$

Combining this with Corollary 1 implies that $T_{YT}(\mathbb{B}=0) \geq T_{YT}(\mathbb{B}=1)$ and $T_P(0) \leq T_P(1)$. □

B Further Tables

Table 3.8: Robustness Check: Varying the Post Cutoff, Extensive Margin

	All Contents			Free Contents			Paid Contents		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$post_t^{Dec2017} \times YT_i$	0.00			0.01*			0.01		
	(0.01)			(0.01)			(0.01)		
$post_t^{Jan2018} \times YT_i$		0.01			0.01			0.01**	
		(0.01)			(0.01)			(0.01)	
$post_t^{Feb2018} \times YT_i$			0.01			0.01*			0.02**
			(0.01)			(0.01)			(0.01)
User FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Post Cutoff	Dec 17	Jan 18	Feb 18	Dec 17	Jan 18	Feb 18	Dec 17	Jan 18	Feb 18
Mean	.533	.533	.533	.32	.32	.32	.403	.403	.403
R^2	.613	.613	.613	.529	.529	.529	.64	.64	.64
N	108,315	108,315	108,315	108,315	108,315	108,315	108,315	108,315	108,315

Notes: Standard errors clustered by CC are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.9: Robustness Check: Varying the Post Cutoff, Intensive Margin

	All Contents			Free Contents			Paid Contents		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$post_t^{Dec2017} \times YT_i$	0.08***			-0.02			0.07***		
	(0.02)			(0.03)			(0.03)		
$post_t^{Jan2018} \times YT_i$		0.07***			-0.02			0.06**	
		(0.02)			(0.03)			(0.03)	
$post_t^{Feb2018} \times YT_i$			0.07***			0.01			0.05**
			(0.02)			(0.03)			(0.02)
User FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Post Cutoff	Dec 17	Jan 18	Feb 18	Dec 17	Jan 18	Feb 18	Dec 17	Jan 18	Feb 18
Mean	1.39	1.39	1.39	.929	.929	.929	1.24	1.24	1.24
R^2	.716	.716	.716	.661	.661	.661	.704	.704	.704
N	57,006	57,006	57,006	33,394	33,394	33,394	42,886	42,886	42,886

Notes: Standard errors clustered by CC are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

C Event Studies

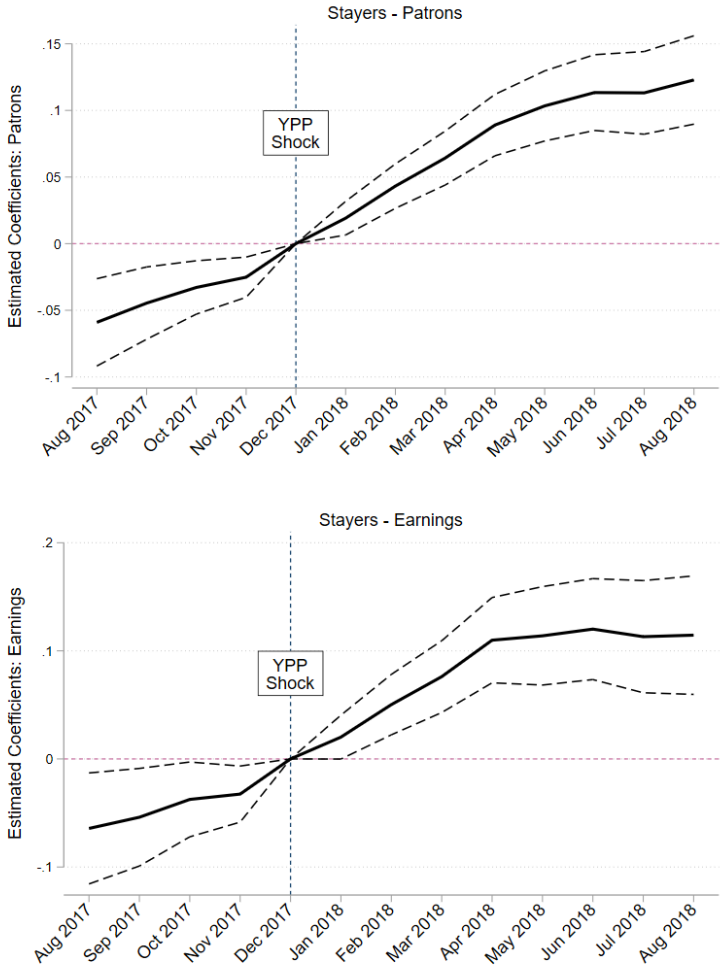


Figure 3.4: Event Study: Log of Number of Patrons and Earnings

Notes: In line with Equation 3.5.2, the log of $Patrons_{it}$, $Earnings_{it}$ are regressed on CC fixed effects; time fixed effects, and on the products between YT_i and a full set of dummy variables for each month. The graphs plot the estimated coefficients on these products. The value of the coefficient corresponding to December 2017 is normalized to zero. The sample includes months between August 2017 and August 2018. Standard errors (5%) are clustered by user.

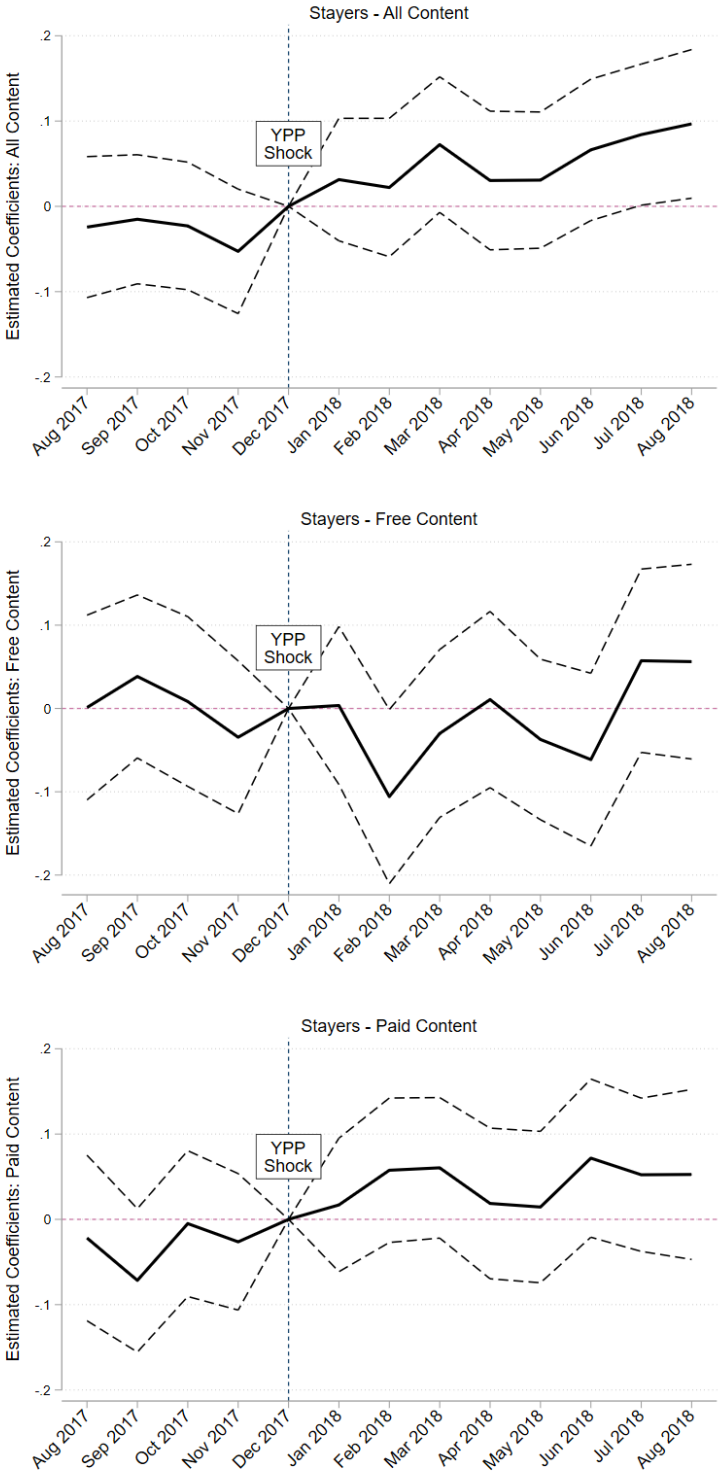


Figure 3.5: Event Study: Log of Number of Contents

Notes: In line with Equation 3.5.2, the log of $Content_{it}$, $Content_{it}^{Free}$, $Content_{it}^{Paid}$ are regressed on CC fixed effects; time fixed effects, and on the products between YT_i and a full set of dummy variables for each month. The graphs plot the estimated coefficients on these products. The value of the coefficient corresponding to December 2017 is normalized to zero. The sample includes months between August 2017 and August 2018. Standard errors (5%) are clustered by user.

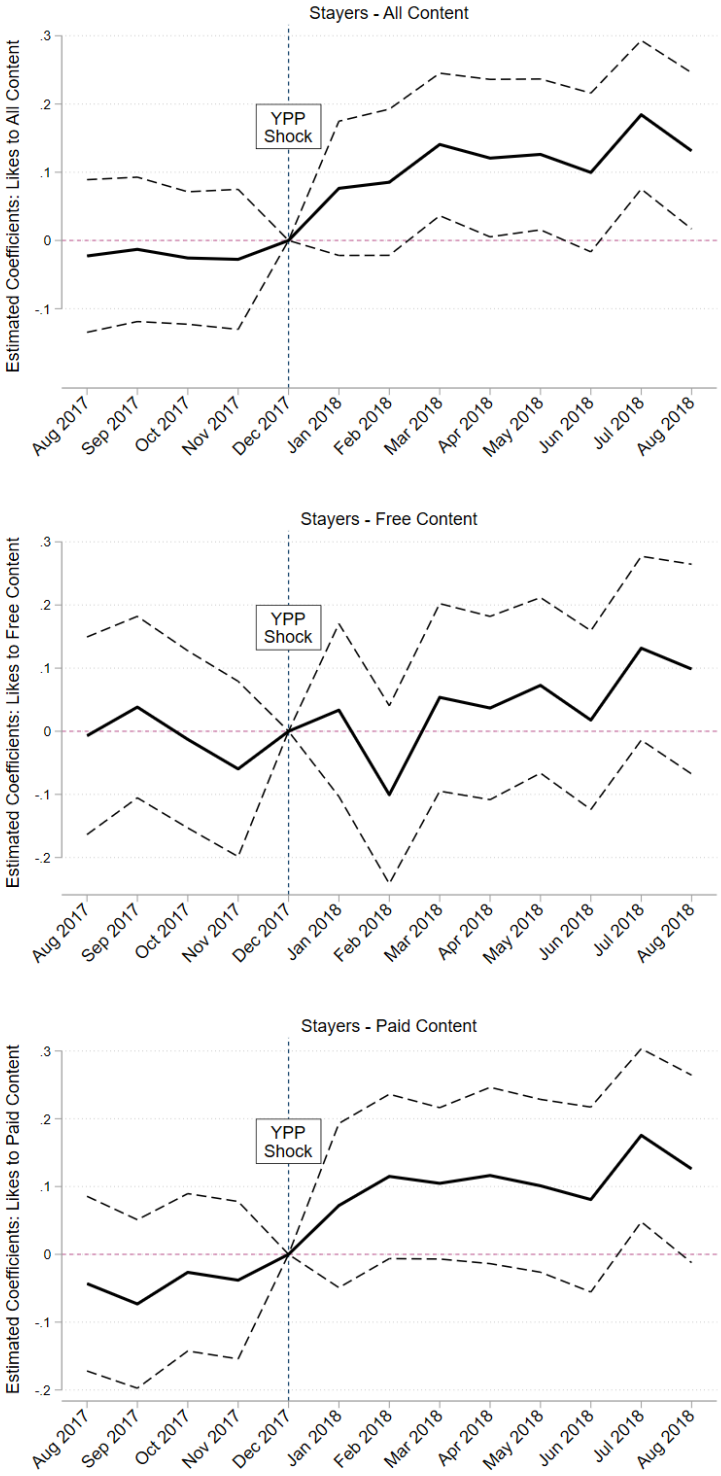


Figure 3.6: Event Study: Log of Number of Likes of Contents

Notes: In line with Equation 3.5.2, the log of $Likes_{it}$, $Likes_{it}^{Free}$, $Likes_{it}^{Paid}$ are regressed on CC fixed effects; time fixed effects, and on the products between YT_i and a full set of dummy variables for each month. The graphs plot the estimated coefficients on these products. The value of the coefficient corresponding to December 2017 is normalized to zero. The sample includes months between August 2017 and August 2018. Standard errors (5%) are clustered by user.

Chapter 4

Conclusion

While offering numerous opportunities to foster innovations, economic efficiency and social connectivity, the digital economy also creates new challenges for society and policy. To mitigate some of these challanges, such as technologies as market entry barriers and online hate speech, policy makers and platform managers have taken interventions in the form of public support schemes, regulation, and platform specific house rules. This doctoral thesis empirically investigates three examples of these interventions and analyses their effectiveness as well as potential side effects. It contributes to the understanding of how interventions in the digital economy influence firms' and individuals' behaviour and what characteristics should be considered when designing those interventions.

The first chapter uncovers an unintended policy effect of a public support scheme for firms in economically lagging regions in Germany. It exploits variation in the eligibility and the size of potential investment subsidies of firms in order to analyse the relationship with the firms' propensity to adopt cloud services. The empirical results demonstrate that the higher the potential subsidy for investments, the lower the incentive for firms to adopt cloud services. This finding should be considered an unintended policy effect which is highly relevant for policy makers, as the adoption of cloud services has been linked to various benefits on the firm as well as on the aggregate economy level.

The second chapter analyses the pioneering online hate speech regulation on the social media platform Twitter (now X). It exploits the regulation, the German Network Enforcement Act, in a quasi-experimental approach to measure the causal impact of the law on the prevalence of hateful content in a target group of the German-speaking segment of Twitter. The results imply a significant and robust decrease in the intensity and volume of hate speech in posts addressing sensitive migration and religion related topics. Importantly, posts tackling other topics as well as the posting style of users are not affected by the regulation, which is in line with its aim. This chapter highlights that legislation for combating harmful online content can significantly reduce the prevalence of hate speech and contributes to understanding the perspective implications of the European Digital Services Act.

The third chapter investigates mechanisms of the Creator Economy, which has experienced a large growth due to digital platforms. On platforms like YouTube and Patreon, content creators can easily diffuse their content and share it with a large crowd, oftentimes for low or zero pecuniary costs. However, the multi-sided business models of those platforms imply complex

dynamics, as they need to match the needs of different stakeholders, such as content creators, users, and (sometimes) advertisers. This becomes apparent in the example of the YouTube "Adpocalypse" in 2017, when major advertisers fled YouTube due to concerns about their ads appearing alongside objectionable content. This chapter exploits YouTube's subsequent content moderation efforts following the Adpocalypse to measure the cross-platform responses of content creators and content consumers on Patreon. Focusing on content creators that multi-home on YouTube and Patreon, the theoretical model and empirical evidence of this chapter confirm that affected content creators respond strategically and shift their efforts toward Patreon. As a result, consumers also increase their use of Patreon through memberships, comments, and likes. However, we also find that YouTube's content moderation and the shift by content creators and consumers that follows, results in an increase in overall toxicity on Patreon. These findings indicate that governance rules of a single platform cannot be considered in isolation. Instead, platform managers and policy makers need to consider potential cross-platform behavioural adjustments of all stakeholders.

The three chapters of this doctoral thesis demonstrate that interventions in the digital economy can be effective in shaping online discussions, but digital particularities such as rapidly changing technologies and substitution effects need to be carefully considered.

Bibliography

Alecke, Björn and Timo Mitze (2023). "Institutional reforms and the employment effects of spatially targeted investment grants: The case of Germany's GRW". *arXiv:2302.11376*.

Ali, Shiza, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini (2021). "Understanding the Effect of Deplatforming on Social Networks". In: *13th ACM Web Science Conference 2021*, pp. 187–195.

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow (2020). "The welfare effects of social media". *American Economic Review*, 110(3): 629–76.

Allcott, Hunt and Matthew Gentzkow (2017). "Social media and fake news in the 2016 election". *Journal of economic perspectives*, 31(2): 211–236.

Andres, Raphaela and Olga Slivko (2021a). "Combating online hate speech: The impact of legislation on Twitter". *ZEW-Centre for European Economic Research Discussion Paper*, 21(103).

— (2021b). "Regulation of Hate Speech and Hatefulness on German Twitter". *ICIS Proceedings*, 11.

Andrews, Dan, Chiara Criscuolo, and Peter Gal (2016). "The best versus the rest: the global productivity slowdown, divergence across firms and the role of public policy". *OECD Productivity Working Papers*, 2016(5).

Aral, Sinan (2020). *The Hype Machine: How Social Media Disrupts Our Elections, Our Economy, and Our Health–and How We Must Adapt*. Crown. ISBN: 9780525574514. URL: https://books.google.nl/books?id=OO-NEAAAQBAJ.

Armbrust, Michael, Armando Fox, Rean Griffith, Anthony D Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, and Ion Stoica (2010). "A view of cloud computing". *Communications of the ACM*, 53(4): 50–58.

Armstrong, Mark (2006). "Competition in two-sided markets". *The RAND journal of economics*, 37(3): 668–691.

BAFA (2019). *Federal Office of Economic Affairs and Export Control*. URL: https://www.bafa.de/EN/Home/home_node.html.

Beknazar-Yuzbashev, George, Rafael Jiménez Durán, Jesse McCrosky, and Mateusz Stalinski (2022). "Toxic Content and User Engagement on Social Media: Evidence from a Field Experiment". *Available at SSRN 4307346*.

Bertschek, Irene and Reinhold Kesler (2022). "Let the user speak: Is feedback on Facebook a source of firms' innovation?" *Information Economics and Policy*, 60: 100991.

Bhargava, Hemant K (2022). "The creator economy: Managing ecosystem supply, revenue sharing, and platform design". *Management Science*, 68(7): 5233–5251.

— (2023). "If It's Enraging, it's Engaging - Exploitive Design in Information Platforms and the Attention Economy".

BMJV (2020). *Bericht der Bundesregierung zur Evaluierung des Gesetzes zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz – NetzDG)*. Tech. rep. Bundesministerium der Justiz und für Verbraucherschutz.

Borowiecki, Martin, Jon Pareliussen, Daniela Glocker, Eun Jung Kim, Michael Polder, and Iryna Rud (2021). "The impact of digitalisation on productivity: Firm-level evidence from the Netherlands". *OECD Economics Department Working Papers*, 2021(1680).

Borwankar, Sameer, Jinyang Zheng, and Karthik Natarajan Kannan (2022). "Democratization of Misinformation Monitoring: The Impact of Twitter's Birdwatch Program". *Available at SSRN 4236756*.

Boudreau, Kevin J and Andrei Hagiu (2009). "Platform rules: Multi-sided platforms as regulators". *Platforms, markets and innovation*, 1: 163–191.

Braghieri, Luca, Ro'ee Levy, and Alexey Makarin (2022). "Social media and mental health". *American Economic Review*, 112(11): 3660–3693.

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson (2019). "Artificial intelligence and the modern productivity paradox". *The economics of artificial intelligence: An agenda*, 23: 23–57.

Buiten, Miriam C, Alexandre de Streel, and Martin Peitz (2020). "Rethinking liability rules for online hosting platforms". *International Journal of Law and Information Technology*, 28(2): 139–166.

Bursztyn, Leonardo, Georgy Egorov, Ruben Enikolopov, and Maria Petrova (2019). "Social Media and Xenophobia: Evidence from Russia". *Communication & Identity eJournal*.

Byrne, David, Carol Corrado, and Daniel Sichel (2021). "The rise of cloud computing: Minding your Ps, Qs and Ks". In: *Measuring and accounting for innovation in the twenty-first century*. NBER Chapters. National Bureau of Economic Research, Inc, pp. 519–551.

Caiani, Manuela and Linda Parenti (2013). "Extreme right groups and the Internet: Construction of identity and source of mobilization". *European and American Extreme Right Groups and the Internet, Londres, Ashgate*: 83–112.

Calvino, Flavio and Luca Fontanelli (2023). "A portrait of AI adopters across countries: Firm characteristics, assets' complementarities and productivity". *OECD Science, Technology and Industry Working Papers*, 2023(02).

Calvino, Flavio, Lea Samek, Mariagrazia Squicciarini, and Cody Morris (2022). "Identifying and characterising AI adopters: A novel approach based on big data". *OECD Science, Technology and Industry Working Papers*, 2022(06).

Casner, Ben and Tat-How Teh (2022). "Content-hosting platforms: discovery, membership, or both?" *SSRN Working Paper 4321577*.

Chandrasekharan, Eshwar, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert (2017). "You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech". *Proceedings of the ACM on Human-Computer Interaction*(, 1. CSCW): 1–22.

Chevalier, Judith A and Dina Mayzlin (2006). "The effect of word of mouth on sales: Online book reviews". *Journal of marketing research*, 43(3): 345–354.

Cho, Jaehan, Timothy DeStefano, Hanhin Kim, Inchul Kim, and Jin Hyun Paik (2023). "What's driving the diffusion of next-generation digital technologies?" *Technovation*, 119: 102477.

Columbus, Louis (2013). "Making Cloud Computing Pay". *Forbes Article*. URL: `https://www.forbes.com/sites/louiscolumbus/2013/04/10/making-cloud-computing-pay-2/?sh=58e472525656`.

Coyle, Diane and David Nguyen (2018). *Cloud computing and national accounting*. Economic Statistics Centre of Excellence (ESCoE) Discussion Papers ESCoE DP-2018-19. Economic Statistics Centre of Excellence (ESCoE).

Criscuolo, Chiara, Ralf Martin, Henry G Overman, and John Van Reenen (2019). "Some causal effects of an industrial policy". *American Economic Review*, 109(1): 48–85.

Cui, Ruomeng, Santiago Gallino, Antonio Moreno, and Dennis J Zhang (2018). "The operational value of social media information". *Production and Operations Management*, 27(10): 1749–1769.

Cusumano, Michael A, Annabelle Gawer, and David B Yoffie (2019). *The business of platforms: Strategy in the age of digital competition, innovation, and power*. Vol. 320. Harper Business New York.

DeStefano, Timothy, Koen De Backer, and Laurent Moussiegt (2017). "Determinants of digital technology use by companies". *OECD Science, Technology and Industry Policy Papers*, 2017(40).

DeStefano, Timothy, Richard Kneller, and Jonathan Timmis (2020). "Cloud computing and firm growth". *CESifo Working Paper Series*, 2020(8306).

Dettmann, Eva, Mirko Titze, and Antje Weyh (2023). "Employment effects of investment grants and firm heterogeneity: Evidence from a staggered treatment adoption approach". *IWH Discussion Papers*, 2023(6).

Deutscher Bundestag (2014). *Unterrichtung durch die Bundesregierung - Koordinierungsrahmen der Gemeinschaftsaufgabe „Verbesserung der regionalen Wirtschaftsstruktur" ab 1. Juli 2014*. Drucksache 18/2200. Berlin. URL: http://dipbt.bundestag.de/dip21/btd/18/022/1802200.pdf.

— (2016). *Regionalpolitischer Bericht der Bund-Länder-Gemeinschaftsaufgabe „Verbesserung der regionalen Wirtschaftsstruktur"*. Drucksache 18/2200. Berlin. URL: https://dserver.bundestag.de/btd/18/075/1807500.pdf.

Duso, Tomaso and Alexander Schiersch (2022). "Let's Switch to the Cloud: Cloud Adaption and Its Effect on IT Investment and Productivity". *DIW Berlin Discussion Paper*, 2022(2017).

Echikson, William and Olivia Knodt (2018). "Germany's NetzDG: A key test for combatting online hate". *CEPS Policy Insight, SSRN 3300636*.

El Sanyoura, Lana and Ashton Anderson (2022). "Quantifying the Creator Economy: A Large-Scale Analysis of Patreon". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 16, pp. 829–840.

ElSherief, Mai, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding (2018). "Hate lingo: A target-based linguistic analysis of hate speech in social media". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 12. 1.

Enikolopov, Ruben, Alexey Makarin, and Maria Petrova (2020). "Social media and protest participation: Evidence from Russia". *Econometrica*, 88(4): 1479–1514.

Ershov, Daniel and Matthew Mitchell (2020). "The effects of influencer advertising disclosure regulations: Evidence from instagram". In: *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 73–74.

Ershov, Daniel and Juan S Morales (2021). "Sharing news left and right: The effects of policies targeting misinformation on social media". *Collegio Carlo Alberto Notebooks*, 2021(651).

Eurostat (2018). *Cloud Computing - Statistics on the Use by Enterprises*. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Cloud_computing_-_statistics_on_the_use_by_enterprises%5C#Enterprises.E2.80.99_dependence_on_cloud_computing.

Farronato, Chiara and Andrey Fradkin (2022). "The welfare effects of peer entry: the case of Airbnb and the accommodation industry". *American Economic Review*, 112(6): 1782–1817.

Feher, Adam (2023). "How to enforce platforms' liability?" *Working paper*.

Florida, Richard (2022). *The rise of the creator economy*. Tech. rep. The Creative Class Group.

Fortuna, Paula, Juan Soler, and Leo Wanner (2020). "Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of hate speech datasets". In: *Proceedings of the 12th language resources and evaluation conference*, pp. 6786–6794.

Gal, Peter, Giuseppe Nicoletti, Theodore Renault, Stéphane Sorbe, and Christina Timiliotis (2019). "Digitalisation and productivity: In search of the holy grail – Firm-level empirical evidence from EU countries". *OECD Economics Department Working Papers*, 2019(1533).

Gierten, David, Steffen Viete, Raphaela Andres, and Thomas Niebel (2021). "Firms going digital: Tapping into the potential of data for innovation". *OECD Digital Economy Papers*, 2021(320).

Goh, Khim-Yong, Cheng-Suang Heng, and Zhijie Lin (2013). "Social media brand community and consumer behavior: Quantifying the relative impact of user-and marketer-generated content". *Information systems research*, 24(1): 88–107.

Griffin, Rachel (2021). "New School Speech Regulation and Online Hate Speech: A Case Study of Germany's NetzDG". *Available at SSRN: 3920386*.

Gu, Bin and Qiang Ye (2014). "First step in social media: Measuring the influence of online management responses on customer satisfaction". *Production and Operations Management*, 23(4): 570–582.

Hagiu, Andrei and Julian Wright (2015). "Marketplace or reseller?" *Management Science*, 61(1): 184–203.

— (2019). "Controlling vs. enabling". *Management Science*, 65(2): 577–595.

Halikiopoulou, Daphne (2018). "A right-wing populist momentum? A review of 2017 elections across Europe". *JCMS: Journal of Common Market Studies*, 56(S1): 63–73.

Haller, Stefanie A. and Iulia Siedschlag (2011). "Determinants of ICT adoption: Evidence from firm-level data". *Applied Economics*, 43(26): 3775–3788.

Han, Xiaochuang and Yulia Tsvetkov (2020). "Fortifying toxic speech detectors against veiled toxicity". *arXiv preprint arXiv:2010.03154*.

Haug, Katharina Candel, Tobias Kretschmer, and Thomas Strobel (2016). "Cloud adaptiveness within industry sectors–Measurement and observations". *Telecommunications policy*, 40(4): 291–306.

Hölig, Sascha and Uwe Hasebrink (2020). *Reuters Institute Digital News Report 2020: Ergebnisse für Deutschland*. Hans-Bredow-Institut für Medienforschung an der Universität Hamburg.

Huang, Ni, Yili Hong, and Gordon Burtch (2016). "Social network integration and user content generation: Evidence from natural experiments". *MIS Quarterly*, 17(001).

Jiménez Durán, Rafael (2022). "The economics of content moderation: Theory and experimental evidence from hate speech on Twitter". *Available at SSRN: 4044098*.

Jiménez Durán, Rafael, Karsten Müller, and Carlo Schwarz (2022). "The Effect of Content Moderation on Online and Offline Hate: Evidence from Germany's NetzDG". *Available at SSRN 4230296*.

Jin, Wang and Kristina McElheran (2017). "Economies before scale: survival and performance of young plants in the age of cloud computing". *Rotman School of Management Working Paper*, 2017(3112901).

Kasakowskij, Thomas, Julia Fürst, Jan Fischer, and Kaja J Fietkiewicz (2020). "Network enforcement as denunciation endorsement? A critical study on legal enforcement in social media". *Telematics and Informatics*, 46: 101317.

Ker, Daniel (2021). "Measuring cloud services use by businesses". *OECD Digital Economy Papers*, 2021(304).

Kesler, Reinhold (2022). "The Impact of Apple's App Tracking Transparency on App Monetization". *SSRN 4090786*.

King, Gary and Richard Nielsen (2019). "Why propensity scores should not be used for matching". *Political analysis*, 27(4): 435–454.

El-Komboz, Lena Abou, Anna Kerkhof, and Johannes Loh (2023). "Platform Partnership Programs and Content Supply: Evidence from the YouTube "Adpocalypse"". *CESifo Working Paper*, 2023(10363).

Lau, Raymond Yiu Keung, Wenping Zhang, and Wei Xu (2018). "Parallel aspect-oriented sentiment analysis for sales forecasting with big data". *Production and Operations Management*, 27(10): 1775–1794.

Lefouili, Yassine and Leonardo Madio (2022). "The economics of platform liability". *European Journal of Law and Economics*, 53(3): 319–351.

Li, Hui and Feng Zhu (2021). "Information transparency, multihoming, and platform competition: A natural experiment in the daily deals market". *Management Science*, 67(7): 4384–4407.

Liesching, Marc, Chantal Funke, Alexander Hermann, Christin Kneschke, Carolin Michnic, Linh Nguyen, Johanna Prüßner, Sarah Rudolph, and Vivien Zschammer (2021). *Das NetzDG in der praktischen Anwendung: Eine Teilevaluation des Netzwerkdurchsetzungsgesetzes*. Carl Grossmann Verlag.

Liu, Yi, Pinar Yildirim, and Z John Zhang (2022). "Implications of revenue models and technology for content moderation strategies". *Marketing Science*, 41(4): 831–847.

Luca, Michael (2015). "User-generated content and social media". In: *Handbook of Media Economics*. Vol. 1, pp. 563–592.

Madio, Leonardo and Martin Quinn (2023). "Content moderation and advertising in social media platforms". *SSRN 3551103*.

Maffini, Giorgia, Jing Xing, and Michael P. Devereux (2019). "The impact of investment incentives: Evidence from UK corporation tax returns". *American Economic Journal: Economic Policy*, 11(3): 361–389.

Magids, Scott, Alan Zorfas, and Daniel Leemon (2015). "The new science of customer emotions". *Harvard Business Review*, 76(11): 66–74.

Mallipeddi, Rakesh, Ramkumar Janakiraman, Subodha Kumar, and Seema Gupta (2021). "The effects of social media content created by human brands on engagement: Evidence from indian general election 2014". *Information Systems Research*, 32(1): 212–237.

Marthews, Alex and Catherine E Tucker (2017). "Government surveillance and internet search behavior". *SSRN 2412564*.

Mell, Peter M. and Timothy Grance (2011). *SP 800-145. The NIST definition of cloud computing*. National Institute of Standards & Technology. Gaithersburg, MD, United States. URL: https://csrc.nist.gov/publications/detail/sp/800-145/final.

Mondal, Mainack, Leandro Araújo Silva, and Fabrıécio Benevenuto (2017). "A measurement study of hate speech in social media". *Proceedings of the 28th ACM conference on hypertext and social media*: 85–94.

Müller, Karsten and Carlo Schwarz (2021). "Fanning the flames of hate: Social media and hate crime". *Journal of the European Economic Association*, 19(4): 2131–2167.

— (2023). "From hashtag to hate crime: Twitter and antiminority sentiment". *American Economic Journal: Applied Economics*, 15(3): 270–312.

OECD (2014). "Cloud Computing: The Concept, Impacts and the Role of Government Policy". *OECD Digital Economy Papers*, 240.

— (2015). *OECD Digital Economy Outlook 2015*. OECD Publishing, Paris. URL: https://doi.org/10.1787/9789264232440-en.

Olteanu, Alexandra, Carlos Castillo, Jeremy Boy, and Kush Varshney (2018). "The effect of extremist violence on hateful speech online". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 12. 1.

Paridar, Mahsa, Mina Ameri, and Elisabeth Honka (2023). "More, Faster, and Better? Effects of Rewards on Incentivizing the Creation of User-Generated Content". *SSRN 4580732*.

Parker, Geoffrey G and Marshall W Van Alstyne (2005). "Two-sided network effects: A theory of information product design". *Management science*, 51(10): 1494–1504.

Pentina, Iryna and Monideepa Tarafdar (2014). "From "information" to "knowing": Exploring the role of social media in contemporary news consumption". *Computers in Human Behavior*, 35: 211–223.

Rauchfleisch, Adrian and Jonas Kaiser (2021). "Deplatforming the far-right: An analysis of YouTube and BitChute". *SSRN 3867818*.

Rochet, Jean-Charles and Jean Tirole (2003). "Platform competition in two-sided markets". *Journal of the European Economic Association*, 1(4): 990–1029.

Schivardi, Fabiano and Tom Schmitz (2020). "The IT revolution and southern Europe's two lost decades". *Journal of the European Economic Association*, 18(5): 2441–2486.

Sen, Ananya, Tom Grad, Pedro Ferreira, and Jörg Claussen (2023). "(How) Does User-Generated Content Impact Content Generated by Professionals? Evidence from Local News". *Management Science*, 0(0).

Siegloch, Sebastian, Nils Wehrhöfer, and Tobias Etzel (2022). "Spillover, efficiency and equity effects of regional firm subsidies". *ECONtribute Discussion Papers Series*, 2022(210).

Srinivasan, Kumar Bhargav, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan (2019). "Content removal as a moderation strategy: Compliance and other outcomes in the ChangeMyView community". *Proceedings of the ACM on Human-Computer Interaction*, 2019(3): 1–21.

Statista (2022). *Cloud Computing Market Size in Europe from 2016 to 2027, by Segment*. Accessed 29th November 2022. URL: https://www.statista.com/forecasts/1235161/europe-cloud-computing-market-size-by-segment.

Teh, Tat-How (2022). "Platform governance". *American Economic Journal: Microeconomics*, 14(3): 213–254.

Tworek, Heidy and Paddy Leerssen (2019). "An analysis of Germany's NetzDG law". *Transatlantic Working Group*.

Uyheng, Joshua and Kathleen M Carley (2021). "Characterizing network dynamics of online hate communities around the COVID-19 pandemic". *Applied Network Science*, 6(1): 1–21.

Vosoughi, Soroush, Deb Roy, and Sinan Aral (2018). "The spread of true and false news online". *science*, 359(6380): 1146–1151.

Wang, Shuting, Min-Seok Pang, and Paul A Pavlou (2021). "Cure or poison? Identity verification and the posting of fake news on social media". *Journal of Management Information Systems*, 38(4): 1011–1038.

Wlömert, Nils, Dominik Papies, Michel Clement, and Martin Spann (2024). "Frontiers: The interplay of user-generated content, content industry revenues, and platform regulation: Quasi-experimental evidence from YouTube". *Marketing Science*, 43(1): 1–12.

Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov (2020). "Political effects of the internet and social media". *Annual review of economics*, 12: 415–438.

Zolas, Nikolas, Zachary Kroff, Erik Brynjolfsson, Kristina McElheran, David N Beede, Cathy Buffington, Nathan Goldschlag, Lucia Foster, and Emin Dinlersoz (2020). "Advanced Technologies Adoption and Use by U.S. Firms: Evidence from the Annual Business Survey". *NBER Working Paper Series*, 2020(28290).

Résumé en francais

Cette thèse étudie de manière empirique les interventions politiques et managériales dans l'économie numérique et analyse leur efficacité et leurs effets secondaires potentiels.

Le premier chapitre met en évidence un effet politique involontaire d'un régime d'aide publique aux entreprises, mis en place dans les régions allemandes ayant un retard sur le plan économique. Aujourd'hui, les entreprises sont confrontées à la décision d'investir dans une infrastructure informatique sur site ou d'acheter des services numériques de pointe. Bien que l'adoption des services du *Cloud* ait été associée à divers bienfaits pour l'entreprise et pour l'économie dans son ensemble, tels que la facilitation du développement d'innovations en matière de produits et de processus et l'augmentation de la concurrence sur le marché et, les programmes d'aide publique se limitent encore souvent à une aide à l'investissement. Ce chapitre analyse donc si les entreprises qui peuvent bénéficier d'aides à l'investissement sont dissuadées d'utiliser des services d'informatique du Cloud. Nous exploitons des variations dans les conditions d'éligibilité et dans les sommes des potentielles subventions, afin d'analyser leur influence sur la propension des entreprises à adopter des services du Cloud.
Les résultats empiriques montrent que les entreprises qui ont accès à des aides à l'investissement sont moins enclines à adopter les services du Cloud. Conformément à ce résultat, nous montrons que plus la subvention potentielle à l'investissement est élevée, moins les entreprises sont incitées à adopter des services du Cloud. Cet effet involontaire de la politique est très pertinent pour les décideurs politiques, car l'adoption des services du Cloud a été associée à divers bienfaits pour

l'entreprise, ainsi que pour l'économie dans son ensemble.

Le deuxième chapitre analyse une réglementation pionnière en matière de discours haineux en ligne sur la plateforme de médias sociaux X (ex Twitter). Nous y exploitons l'introduction d'une régulation en Allemagne, qui s'applique aux réseaux sociaux, dans une approche quasi-expérimentale pour mesurer l'impact causal de la loi sur la prévalence des contenus haineux dans un groupe cible du segment germanophone de X. Les résultats impliquent une diminution significative et robuste de l'intensité des discours de haine dans les tweets abordant des sujets sensibles liés à l'immigration et à la religion. En plus, le volume de tweets haineux a également diminué après la mise en œuvre de la loi. Il est important de noter que les tweets abordant d'autres sujets, ainsi que le style de tweet des utilisateurs ne sont pas affectés par le règlement, ce qui est conforme à son objectif. En outre, si nous constatons que les tweets haineux suscitent un engagement plus élevé de la part des utilisateurs en général, il n'y a pas d'augmentation supplémentaire du nombre de likes, de commentaires ou de partages des tweets haineux restants après la mise en place de la réglementation. Ce chapitre montre que la législation visant à lutter contre les contenus en ligne préjudiciables peut réduire de manière significative la prévalence du discours de haine. Par ailleurs, il contribue à comprendre les potentiels effets que pourrait avoir le Digital Service Act, instauré par l'Union Européenne.

Le troisième chapitre étudie les mécanismes de l'«économie des créateurs». Elle s'appuie sur un modèle commercial de plateforme multi-face, mettant en relation les créateurs de contenu, les utilisateurs et (parfois) les annonceurs. Répondre aux besoins de ces différents partis est un défi complexe, comme le montre l'impact de l'«Adpocalypse» de YouTube en 2017. Lors de cet épisode, de grands annonceurs ont quitté YouTube parce qu'ils craignaient que leurs publicités soit associées à des contenus répréhensibles. Ce chapitre exploite les efforts de modération de contenu déployés par YouTube à la suite de l'Adpocalypse pour mesurer les réponses multiplateformes des créateurs et des consommateurs de contenu sur Patreon.
Le modèle théorique et les données empiriques de ce chapitre confirment que les créateurs de contenu qui pratiquent le multi-homing sur YouTube et Patreon réagissent de manière stratégique et concentrent leurs efforts sur Patreon. Par conséquent, les consommateurs augmentent leur utilisation de Patreon par le biais d'adhésions, de commentaires et d'appréciations. Toutefois,

nous constatons également que la modération du contenu de YouTube et le changement de comportement des créateurs et des consommateurs qui en découle entraînent une augmentation de la toxicité sur Patreon. Ces résultats indiquent que les règles de gouvernance d'une seule plateforme ne peuvent pas être considérées de manière isolée. Au contraire, les gestionnaires de plateformes et les décideurs politiques doivent prendre en compte les réactions potentielles de toutes les parties prenantes.

Les trois chapitres de cette thèse démontrent que les interventions dans l'économie numérique peuvent être efficaces pour façonner les discussions en ligne, mais les particularités numériques telles que l'évolution rapide des technologies et les effets de substitution doivent être soigneusement prises en compte.

ECOLE
DOCTORALE

Titre : Interventions politiques dans l'économie numérique: Essais empiriques

Mots clés : plateformes numériques, politique publique, contenu généré par les utilisateurs

Résumé : Cette thèse étudie de manière empirique les interventions politiques et managériales dans l'économie numérique et analyse leur efficacité et leurs effets secondaires potentiels. Le premier chapitre met en évidence un effet politique involontaire d'un régime d'aide publique aux entreprises, mis en place dans les régions allemandes ayant un retard sur le plan économique. Nous exploitons des variations dans les conditions d'éligibilité et dans les sommes des potentielles subventions, afin d'analyser leur influence sur la propension des entreprises à adopter des services du Cloud. Les résultats empiriques montrent que plus la subvention potentielle à l'investissement est élevée, moins les entreprises sont incitées à adopter des services du Cloud. Cet effet involontaire de la politique est très pertinent pour les décideurs politiques, car l'adoption des services du Cloud a été associée à divers bienfaits pour l'entreprise, ainsi que pour l'économie dans son ensemble. Le deuxième chapitre analyse une réglementation pionnière en matière de discours haineux en ligne sur la plateforme de médias sociaux Twitter (maintenant X). Nous y exploitons l'introduction d'une régulation en Allemagne, qui s'applique aux réseaux sociaux, dans une approche quasi-expérimentale pour mesurer l'impact causal de la loi sur la prévalence des contenus haineux dans un groupe cible du segment germanophone de Twitter. Les résultats impliquent une diminution significative et robuste de l'intensité et du volume des discours de haine dans les tweets abordant des sujets sensibles liés à l'immigration et à la religion. Il est important de noter que les tweets abordant d'autres sujets, ainsi que le style de tweet des utilisateurs ne sont pas affectés par le règlement, ce qui est conforme à son objectif. Ce chapitre montre que la législation visant à lutter contre les contenus en ligne préjudiciables peut réduire de manière significative la prévalence du discours de haine. Par ailleurs, il contribue à comprendre les potentiels effets que pourrait avoir le Digital Service Act, instauré par l'Union Européenne. Le troisième chapitre étudie les mécanismes de l'≪économie des créateurs≫. Elle s'appuie sur un modèle commercial de plateforme multi-face, mettant en relation les créateurs de contenu, les utilisateurs et (parfois) les annonceurs. Répondre aux besoins de ces différents partis est un défi complexe, comme le montre l'impact de l'≪Adpocalypse≫ de YouTube en 2017. Lors de cet épisode, de grands annonceurs ont quitté YouTube parce qu'ils craignaient que leurs publicités soit associées à des contenus répréhensibles. Ce chapitre exploite les efforts de modération de contenu déployés par YouTube à la suite de l'Adpocalypse pour mesurer les réponses multiplateformes des créateurs et des consommateurs de contenu sur Patreon. Le modèle théorique et les données empiriques de ce chapitre confirment que les créateurs de contenu qui pratiquent le multi-homing sur YouTube et Patreon réagissent de manière stratégique et concentrent leurs efforts sur Patreon.Par conséquent, les consommateurs augmentent leur utilisation de Patreon par le biais d'adhésions, de commentaires et d'appréciations. Toutefois, nous constatons également que la modération du contenu de YouTube et le changement de comportement des créateurs et des consommateurs qui en découle entraînent une augmentation de la toxicité sur Patreon. Ces résultats indiquent que les règles de gouvernance d'une seule plateforme ne peuvent pas être considérées de manière isolée. Au contraire, les gestionnaires de plateformes et les décideurs politiques doivent prendre en compte les réactions potentielles de toutes les parties prenantes. Les trois chapitres de cette thèse démontrent que les interventions dans l'économie numérique peuvent être efficaces pour façonner les discussions en ligne, mais les particularités numériques telles que l'évolution rapide des technologies et les effets de substitution doivent être soigneusement prises en compte.

ECOLE DOCTORALE

Title : Empirical Essays on Policy Interventions in the Digital Economy

Keywords : digital platforms, public policy, user-generated content

Abstract : This thesis empirically investigates policy and managerial interventions in the digital economy and analyses their effectiveness and potential side effects. The first chapter uncovers an unintended policy effect of a public support scheme for firms in economically lagging regions in Germany. It exploits variation in the eligibility and the size of potential investment subsidies of firms in order to analyse the relationship with the firms' propensity to adopt cloud services. The empirical results demonstrate that the higher the potential subsidy for investments, the lower the incentive for firms to adopt cloud services. This finding should be considered an unintended policy effect which is highly relevant for policy makers, as the adoption of cloud services has been linked to various benefits on the firm as well as on the aggregate economy level.
The second chapter analyses the pioneering online hate speech regulation on the social media platform Twitter (now X). It exploits the regulation, the German Network Enforcement Act, in a quasi-experimental approach to measure the causal impact of the law on the prevalence of hateful content in a target group of the German-speaking segment of Twitter. The results imply a significant and robust decrease in the intensity and volume of hate speech in posts addressing sensitive migration and religion related topics. Importantly, posts tackling other topics as well as the posting style of users are not affected by the regulation, which is in line with its aim. This chapter highlights that legislation for combating harmful online content can significantly reduce the prevalence of hate speech and contributes to understanding the perspective implications of the European Digital Services Act. The third chapter investigates mechanisms of the Creator Economy, which capitalizes on a multi-sided business model, connecting content creators, users, and (sometimes) advertisers. Matching the needs of these different stakeholders is a complex challenge, as evidenced by the impact of the YouTube "Adpocalypse" in 2017, when major advertisers fled YouTube due to concerns about their ads appearing alongside objectionable content. This chapter exploits YouTube's subsequent content moderation efforts following the Adpocalypse to measure the cross-platform responses of content creators and content consumers on Patreon. Focusing on content creators that multi-home on YouTube and Patreon, the theoretical model and empirical evidence of this chapter confirm that these content creators respond strategically and shift their efforts toward Patreon. As a result, consumers also increase their use of Patreon through memberships, comments, and likes. However, we also find that YouTube's content moderation and the shift by content creators and consumers that follows, results in an increase in toxicity on Patreon. These findings indicate that governance rules of a single platform cannot be looked at in isolation. Instead, platform managers and policy makers need to consider potential cross-platform reactions of all stakeholders. The three chapters of this thesis demonstrate that interventions in the digital economy can be effective in shaping online discussions, but digital particularities such as rapidly changing technologies and substitution effects need to be carefully considered.

Institut Polytechnique de Paris
91120 Palaiseau, France