FELIPE MATIAS BAILEZ VIANA

DEPLATFORMING WHATSAPP

IMPACTS OF TAKING DOWN ANTI-DEMOCRATIC GROUPS

São Paulo
2022

FELIPE MATIAS BAILEZ VIANA

DEPLATFORMING WHATSAPP

IMPACTS OF TAKING DOWN ANTI-DEMOCRATIC GROUPS

Work presented to the School of Economics, Business, Accounting and Actuary of the University of São Paulo to obtain bachelor degree in Economics.

Advisor:
Luis Eduardo Negrão Meloni

São Paulo
2022

I dedicate this work to my parents,
Omar and Ana.

ACKNOWLEDGMENTS

I thank to my parents Omar and Ana for giving me the inspiration and teaching me the importance of humanistic values that guided my decisions until now, and also for the suggestions and critics to this work;

to my brother Tomás and my sister Carol, for all the support and companionship;

to my associates and long-time high school friends, Giancarlo Rocha and Enzo Robaina that accompanied me through the last three years in building this messaging monitor technology that resulted in the creation of Palver, which without them this work wouldn't be possible;

to all crew from Palver: Luis Fakhouri, Eduardo Junqueira, Miguel Lian and Edmar Filho that helped many times during this election and before;

to Vinicius Princiotti and Jamil Civitarese for great suggestions, ideas and critics that improved this work;

to my friends from "Not GEEP" (Mavi, Sofia, Sallum, Lian, Tominhas, Theo, Luca, Gui, Francisco and Leo) that encouraged me to continue this work when my RDD plots still looked like Jackson Pollock's paintings;

to all my teammates from FEA Rugby Club (FRC) that helped me innumerous times with academic and extra-curricular activities and made my college years much more joyful;

to all members from EPEP-USP (Estudos de Política em Pauta) and FEA.dev that unlocked new possibilities for me during my undergrad

to Alexandra Elbakyan, for building Sci-Hub, that helped me in this work and empowers the general-public with access to scientific literature;

to my advisor, Professor Luis Meloni for the support and acceptance in helping me with this work;

to all professors at the Department of Economics from FEA-USP and the teachers I had at Instituto Federal Fluminense that contributed to my studies and learning;

to viola caipira, samba and música popular brasileira that surely kept me alive during the last years.

"The internet, our greatest tool of emancipation, has been transformed into the most dangerous facilitator of totalitarianism we have ever seen. The internet is a threat to human civilization."

-- Julian Assange

RESUMO

O uso das redes sociais podem criar câmaras de eco aumentando a polarização política. Estudos foram conduzidos utilizando dados das principais redes sociais para medir relações causais de moderação de conteúdo e banimento de usuários. O Tribunal Superior Eleitoral (TSE) de Brasil ordenou o banimento de grupos de WhatsApp dentro de um cluster engajado com mobilização anti-democratica. Esta monografia explora discontinuidade utilizando RDiT para medir impactos desta decisão em grupos anti-democráticos que não foram banidos. Apesar de ser a principal rede social no Brasil, literatura analizando dados de WhatsApp ainda é escassa. Resultados sugerem que medida judicial não teve efeito no fluxo de conversa em nenhum tipo de grupo, mas reduziu o alcance das mensagens virais espalhadas por grupos anti-democraticos no WhatsApp.

Palavras-Chave: deplatforimng, RDiT, eleições, redes sociais, WhatsApp.

JEL Codes: C38, C45, D72, L82, Z13

ABSTRACT

The use of social media can produce echo-chamber and increase group polarization. Studies have been conducted using data from mainstream social media to measure causal relationship of deplatforming policy and banning accounts. The Supreme Electoral Court (TSE) from Brazil issued a decision to take down the cluster of WhatsApp groups engaged with anti-democratic activities. This dissertation explores discontinuity using RDiT design to measure impacts of these decision on anti-democratic groups that were not banned. There is scarce literature analyzing WhatsApp data even though it's the biggest social media in Brazil. Findings suggests that judicial decision had no effect on conversation flow on any type of groups, but reduced the reached audience of anti-democratic viral messages spread on WhatsApp.

Keywords: deplatforimng, RDiT, elections, social media, WhatsApp.

JEL Codes: C38, C45, D72, L82, Z13

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

"Eu quero entrar na rede
Promover um debate
Juntar via Internet
Um grupo de tietes de Connecticut"

-- Gilberto Gil

The advent of internet and widespread access to it's technology shaped the world in the last four decades. It improved the capability of citizens to convey and receive information to and from the governments Margetts (2013). Depending on circumstances, internet was a catalyst tool to increase democracy or an instrument for authoritarianism Best and Wade (2009). The sophisticated use of misinformation potentiated by machine-learning methods was applied to hijack democratic process in recent years, the case of Cambdige Analytica being the most iconic example Isaak and Hanna (2018).

With the new challenges to modern democracy, it becomes imperative that government and researches develop technological and regulatory instruments to counter threats to democratic order. One of the main used social media in Brazil is WhatsApp and misinformation travelled through it could put democratic institutions in danger [1]. Shortly after the 2022 elections in Brazil, the Supreme Electoral Court (TSE) issued a decision to shut down many WhatsApp groups engaging with extremist and anti-democratic mobilization. In this context, is essential to find mechanisms that protect democracy without damaging free-speech and individual liberties.

The objective of this work is to contribute to economic literature investigating if judicial interventions on WhatsApp are efficient to reduce viral messages and how it can impact voters communication through these channels. The specific objectives are to evaluate if Supreme Electoral Court decision to ban anti-democratic groups:

1. Suppressed conversation of people using the app in any groups that were not banned;
2. Changed the number of viral forwarded messages through themed groups (anti-democratic and others themes);

[1] https://noticias.uol.com.br/colunas/cristina-tardaguila/2022/09/28/um-terco-das-mensagens-de-whatsapp-sobre-urnas-sao-negativas.htm

3. If there was a change on viral messages, how much it increased or decreased on average.

This work is organized as follow: Chapter 2 presents a literature review in three sections. The first section introduces the importance of information to democracy and the role of social media in recent years. The second section shows some of the works on deplatforming and content moderation and it's impacts so far. The third section gives institutional context to brazilian elections and judicial decision that banned WhatsApp groups. Chapter 3 explains how data was collected, the method applied to classify anti-democratic groups and the transformation to data to create panels used in estimates. Chapter 4 introduces research design used to extract causal effects of this events and the potential bias that could affect estimate. Chapter 5 contains all results obtained and a discussion of it's implications. Chapter 6 concludes the work and open possible next steps on this investigation.

2 LITERATURE REVIEW

2.1 Democracy in the era of social media

Information has an important role in democracy, people will choose to vote for politicians that best represent their interests Przeworski et al. (1999). If information is distorted when transmited to population it can provoke real impact on society. In news outlets, the portrayal of hispanic and black men associated with crime and low-skill jobs perpetuates stereotypes Ash et al. (2021). In an electoral process, it could impact the outcome of election results . The unfavorable television coverage of political debate in Brazilian elections in 1989 made left-wing candidate lose share of voters in the regions with TV broadcast signal Cavgias et al. (2019). Specific cable news television channels can play a role as exogenous shifter of ideological identification of the viewership. Martin and Yurukoglu (2017) used voting data of US Presidential elections and data on cable channels to model evolution of voters ideology. The growth of Fox News on specific regions increased polarization and Republican vote share in presidential elections over time.

With the coming of the internet age, people began to change voting behaviors. Falck et al. (2014) shows that access to high-speed internet had negative impacts on voter turnout in West Germany. They used regional peculiarities of the preexisting telephony network to adress endogeneity in Internet availability. With the widespread use of internet and consumption of information through it's channels, people began to engage in policy debates on social media platforms. As shown by Bessi et al. (2015), a topic extraction strategy was used on Facebook data from Italy and revealed that most consumed contents where related to environment, diet, health and geopolitics. They also modeled user mobility an found that the more active the user, the more likely he was to span on all semantic categories. The continuous use of social networks can produce echo-chambers that increase polarization and impact users behaviors outside internet. Using Facebook municipal level data and anti-refugee hate crimes in Germany, Müller and Schwarz (2017) showed the role of echo chambers and how far-right social media posts can act as propa-

gation mechanism for violent hate crimes.

After the rise of Donald Trump in the 2016 US elections, the use of social media as a political tool got more attention. Müller and Schwarz (2018) constructed an instrument for Twitter using South by Southwest (SXSW) festival that differentiated US counties by early adopters of the micro-blogging platform. They found that since the 2016 presidential primaries, when Donald Trump rises as the Republican party nominee, anti-Muslim hate crimes increased in counties with higher Twitter usage. Using the same instrument, Fujiwara et al. (2021) investigated how Twitter usage affected the election outcomes in the United States. Their findings indicated that Twitter lowered Republican vote share in 2016 and 2020 presidential elections. Exogenous change of rule on Twitter, doubled the maximum limit of characters in a tweet, Jaidka et al. (2019) used natural language processing methods and concluded that it led to less uncivil and more polite discussions online, reducing extremism. A randomized experiment conducted by Levy (2021) subscribed participants to conservative or liberal news outlets on Facebook, his findings suggests that social media algorithms may limit exposure to counter-attitudinal news and increase polarization.

2.2 Deplatforming and content moderation

In January 6 of 2021, far-right extremist stormed the capitol of the United States in an attempt to contest presidential elections results, causing five deaths and others injured[1]. These riots were organized using social media[2] and resulted in the banning of Donald Trump's official twitter account two days after the events. New type of policy enforced by social network companies opened a path to new field of study, weather content moderation and deplatforming are efficient tools against extremism Roberts (2019).

Investigating deplatformed Twitter accounts, Mitts (2021) verifies an increasing of hate-speech on underground platforms like Gab. Analyzing deplatforming of far-right in Youtube, Rauchfleisch and Kaiser (2021) found that removing far-right content was effective in minimizing the reach of disinformation and extreme speech. Using *regression-discontinuity* design (RDD), Freire et al. (2022) measured how content moderation on Youtube affected channel popularity and suggests that moderation may not be effective in changing growth trajectory of channels. Many social media platforms use automated

[1] https://www.npr.org/sections/insurrection-at-the-capitol/2021/01/07/954671745/on-far-right-websites-plans-to-storm-capitol-were-made-in-plain-sight

[2] https://www.nytimes.com/2021/01/06/us/politics/protesters-storm-capitol-hill-building.html

content moderation, Ribeiro et al. (2022) explored Facebook data with fuzzy RD design and found that comment deletion decreased subsequent to rule-breaking rather then hiding content that had insignificant effects. Rogers (2020) highlights that researchers investigating this phenomena, should look with attention to underground social medias. He enquires how Telegram had been used by deplatformed users and if deplatforming in traditional big social media works.

Better understanding how these moderation measures can impact mobilization in real world through social media, specially in dark social[3], is essential to counter extremism and misinformation in modern democracies.

2.3 Institutional context

Brazil's young democracy has only 34 years old, but it has one of the most advanced voting technologies. In 1998 elections, Brazil introduced the use of electronic voting in the national elections. The implementation of this technology wasn't applied to the entire country at once, it used a population cohort to decide which municipalities would start testing the new method. Exploring this discontinuity Fujiwara (2015) shows how positive electronic voting was in reducing residual votes (uncounted and error-ridden) and how it promoted enfranchisement of less educated citizens.

The country held presidential elections in 2022, the two main candidates were left-wing ex-president Lula da Silva (Worker's Party) and far-right incumbent Jair Bolsonaro (Liberal Party). Through his mandate as president, Jair Bolsonaro had many times contested the results of electronic voting[4], even questioning the 2018 results that elected him as president.

In this electoral cycle, one of the main concerns of the Supreme Electoral Court (TSE) was the widespread misinformation about the electoral process. During the campaign, the TSE banned from social media various publications spreading fake news about the electronic voting system, fraud and the judiciary[5]. The media reported how messages in WhatsApp groups were filled with attacks towards the electronic voting system, incited military intervention and defamation of electoral justice Alexandre de Moraes[6], the

[3]Dark social are social networks that content isn't public and can't be measured by traditional analytics methods

[4]https://www.cnnbrasil.com.br/politica/bolsonaro-volta-a-questionar-seguranca-da-urna-eletroni

[5]https://www.tse.jus.br/comunicacao/noticias/2022/Outubro/combate-a-desinformacao-tse-derruba-m

[6]https://oglobo.globo.com/blogs/sonar-a-escuta-das-redes/post/2022/08/mencoes-a-alexandre-de-moraes-crescem-em-grupos-de-whatsapp-em-agosto.ghtml

president of the TSE.

WhatsApp is the most used social network in Brazil[7]. Evangelista and Bruno (2019) suggests that it played important role spreading misinformation in 2018 elections. A better understanding of the dynamics of messaging apps and how they can interfere with political processes is essential to help policy makers regulate them. It also helps to understand the potential dangers of misinformation and polarization in modern democracy. Melo et al. (2021) proposed a method for monitoring data from WhatsApp public groups and generating data. There is still no research using econometric design on this type of data to analyze causal relationship between counter-misinformation strategies and user behaviour on WhatsApp.

October 30 of 2022 ended the second round of brazilian elections resulting in the confirmation of Lula's victory. Shortly after the disclosure, far-right voters started to contest the results. Roads were blocked by truckers and riots were organized in front of military quarters the following day[8]. In the midst of institutional tensions with the protests and anti-democratic mobilization, the Supreme Electoral Court issued an order to WhatsApp and Telegram ordering the removal of the clusters of groups related with "Military Intervention" and asking for a Coup d'Etat[9]. This was the first time the judiciary made a decision to shut down only specific themed groups in WhatsApp. There are still no studies that investigates if this type of intervention is efficient to contain spreading of anti-democratic messages through the groups. Understanding the impacts of this decision can help policy makers design better mechanisms alongside social media platforms to counter the extremist.

[7] https://valorinveste.globo.com/objetivo/gastar-bem/noticia/2021/09/16/80percent-dos-brasileiros-utilizam-o-whatsapp-para-se-comunicar-com-as-marcas-aponta-pesquisa.ghtml

[8] https://www.bloomberg.com/news/articles/2022-11-01/brazil-protests-grow-with-bolsonaro-silent-

[9] https://www1.folha.uol.com.br/poder/2022/11/bolsonaristas-e-tse-fazem-jogo-de-gato-e-rato-com-shtml

3 DATA

3.1 Exploring raw data

The dataset consists in messages extracted from WhatsApp public groups conversations. Data extraction strategy followed similar methodology proposed by Melo et al. (2021) and Resende et al. (2018). WhatsApp invitation links available on indexed web pages[1] were scraped to create a sample of groups. These invitation links were used to join WhatsApp groups and then messages in those phones were stored in a structured database.

New groups were constantly being sampled since November of 2021, Figure 1 shows how groups were added to the sample over time. Complying with local privacy law, no user personal data is stored. By November of 2022, the total of groups were 16.204.

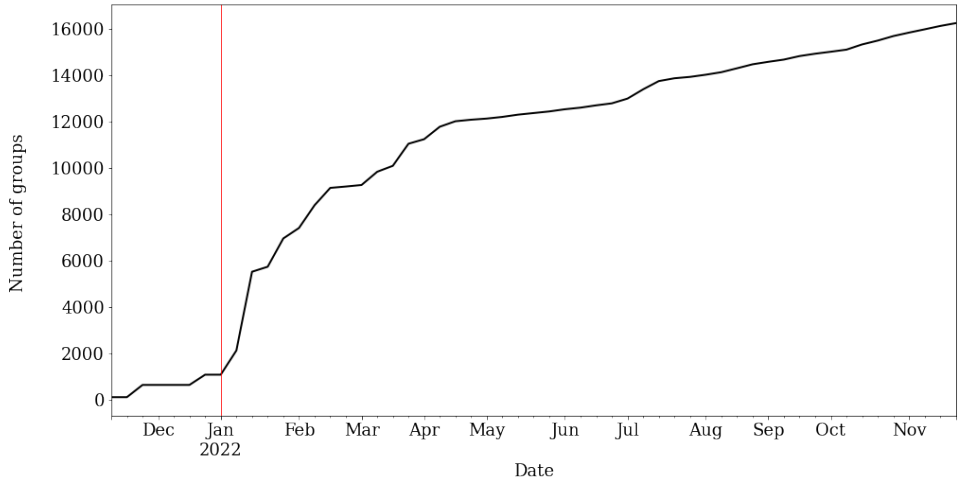

Figure 1: Evolution of quantity of groups on sample over time. The horizontal axis are the dates ranged since November of 2021 until November of 2022. The vertical axis is the number total of groups introduced to the sample at that time. The red line is January 1st of 2022.

[1]Indexed pages are the pages of a website that a search engine has visited and added to its database of web pages.

The data collected is rich but only a some fields will be relevant to achieve the objectives of this research. Selected fields chosen are presented in Table 1. The messages table contains 32 million observations with 16.106 unique groups. In order to analyze the period of the judicial decision, a sample of this data from October and November will be used in this investigation.

Table 1: WhatsApp messages panel data description

Field name	Data type	Description
`_id`	`string`	Unique ID of the message
`chat_id`	`string`	ID of the group where a message was sent
`filename`	`string`	The path where the file was stored
`type_label`	`string`	(textual, document, image, video or audio)
`forwarding_score`	`integer`	Number of times a message was forwarded
`sender_info`	`integer`	Phone area code of message sender
`media_key`	`string`	Unique hash representing a media file
`timestamp`	`integer`	Date and time the message was sent

The following data fields will construct final panel data:

1. `timestamp` - Exact date and time a message was sent by any user;
2. `chat_id` - Identification code representing unique groups;
3. `forwarding_score` - This field carries a numeric value that counts the number of times a message was forwarded. It means that if a user forwards a message from outside the sample, it will carry the number of times it was forwarded before. This value is truncated to a maximum value of 127 that represents the "Forwarded many times" in the app.

To suppress widespread of misinformation, WhatsApp only allows forwarding messages to a total of five conversations each time. When a message reaches "Forwarded many times" status, users can only forward that message to a single conversation[2]. Knowing how forwarding score is computed is important to better understand the impacts of the judicial decision.

[2] https://faq.whatsapp.com/1053543185312573/?helpref=hc_fnav

With the phone area code of the message *sender_info* is possible to know how many individuals are from different regions of Brazil. In the entire sample, a total of is 922.954 users were found in those groups. Figure 2 shows a plot of the percentual share (%) of sampled WhatsApp users by region comparing with brazilian population (IBGE/Censo 2010) by the same regions. This can help shed some light to understand how representative can this sample be.

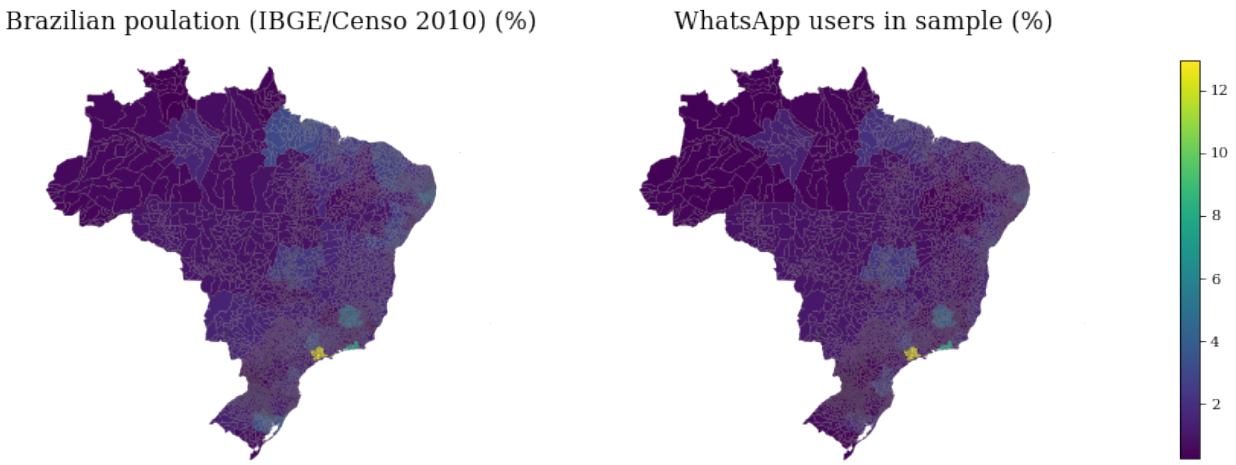

Figure 2: This figure shows a heat-map comparing the regional share of brazilian population according to (Censo-2010) and the sampled users on WhatsApp. The color degrees in the side bar represents the percentage share of each region. Location of users were extracted from the phone area code and account for $N = 922.954$.

It's also interesting to understand if there is any kind of cyclical behaviour of user activity in the app. Taking a randomized sample of only 100.000 messages and then aggregating by 1-minute bins helps visualizing change in message flow through the hours of the day. In Figure 3, a scatter plot is presented and suggest that users are less active during night time.

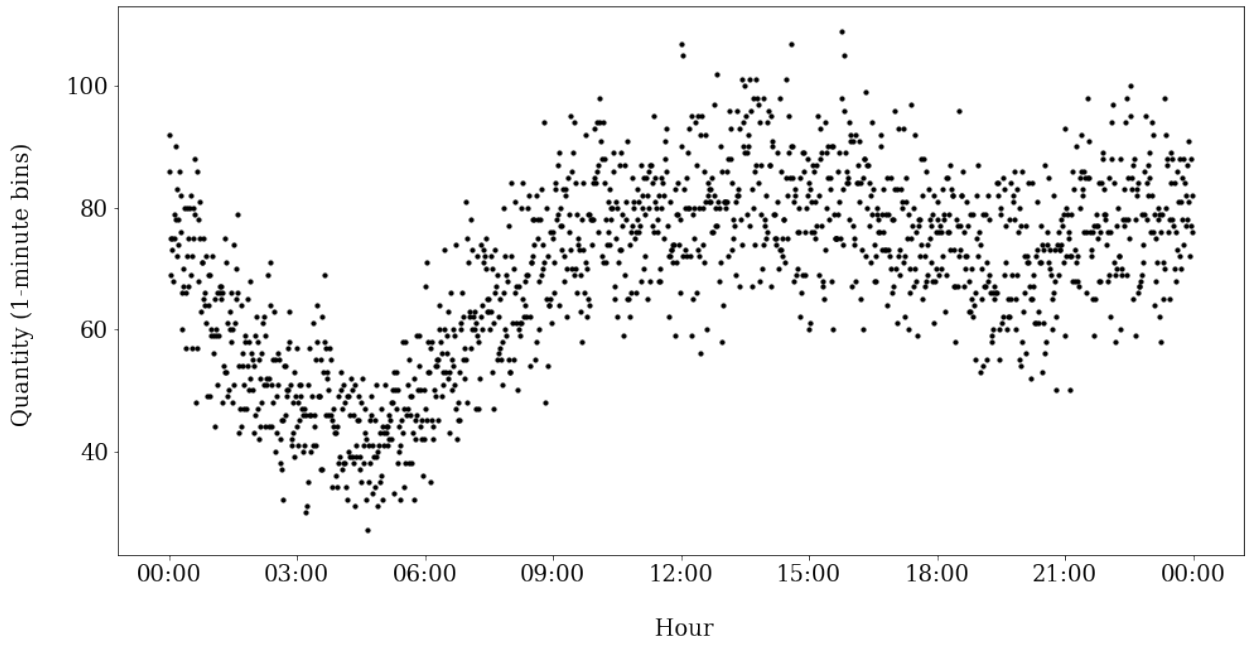

Figure 3: Scatter plot of messaging activity during a day. This plot uses random sample of messages from the entire dataset with $N = 100.000$. Vertical axis shows the quantity of messages sent each minute. Horizontal axis show the hours of an entire day. Each point in this graph represents activity at a minute of a day.

3.2 Classification of anti-democratic groups

Investigating a causal relationship and measuring the outcomes of banning anti-democratic groups requires a method to separate groups by theme. Using Twitter data, Hartung et al. (2017) proposed machine learning method to classify far-right users based on textual clues, traits of emotionality in language use, and linguistic patterns. Lenihan (2021) used network mapping and textual clues to classify users affiliated with far-left political alignment.

Since there are no large public datasets on WhatsApp data, there is still computational method in literature that precisely classifies political alignment of users on messaging apps.

However, Machado et al. (2019) monitored 130 WhatsApp public groups during brazilian elections in 2018 and manually tagged pro-right or pro-left media files on messages in their research. Following this idea, I tagged 105 media files (images, audios, videos and stickers) with high forwarding score (viral) containing far-right content or anti-democratic narrative. Some examples of the images tagged are presented in Figure 4.

(a) Federal Intervention riot convocation

(b) Call for riots in front of military quarters

(c) Asking for military intervention

(d) Federal intervention with Bolsonaro

Figure 4: Examples of anti-democratic images shared on WhatsApp (a,b,c,d)

A window of time containing messages only from October and November was selected. To classify ideological leaning of groups, a filter was applied to keep only observations of WhatsApp groups that were actively sending messages in the days neighboring the beginning and ending of the sampled window. This resulted in a panel with 7.441.087 observations with 5.084 unique groups. The "media key" data field (Table 1) was applied to search for any of the 105 anti-democratic tagged media files in this sample. Of all groups found in this period, 329 unique groups were classified as anti-democratic using tagged media. This produced a separation of 1.700.684 messages in anti-democratic tagged groups and 5.740.403 in all other groups.

3.3 Re-sampling data by time and theme

A new panel was constructed re-sampling data by average 1-minute bins. This granularity was chosen but it can be re-sampled to other frequencies to test robustness of analysis. New data fields separated in treatment (anti-democratic groups), and control (other groups). Those new fields contain four of the dependent variables:

- *treat_q* and *control_q* - Average *quantity* in 1-minute bins of messages sent in treatment or control group;
- *treat_fs* and *control_fs* - Average *forwarding score* in 1-minute bins of messages sent in treatment or control group;

A binary variable named *threshold* was created with value 0 in all periods and 1 after the execution of court decision. A *date* variable was created with the distance from cohort in hours. New data fields presented in Table 2 were created in order to match with model design. Descriptive statistics of the constructed panel are presented in Table 2. Since control group contains 5.084 WhatsApp groups and treatment contains 329, data was normalized to simulate 100 groups, dividing by the total number of groups and multiplying by 100.

Table 2: Summary statistics

	Obs.	Mean	Std. dev.	Min	25%	50%	75%	Max
Quantity of messages								
control_q	76836	1.6	0.8	0	1	1.5	2	11.7
treat_q	76836	6.7	5.3	0	3	6.1	9.4	186
Forwarding score of messages								
control_fs	76836	0.6	3.6	0	0	0.1	0.2	190
treat_fs	76836	28.8	71.5	0	0.3	1.5	39.5	3901.5

Notes:
This table represent summary statistics of the panel that will be used to evaluate this work enquiry. Fields created are quantity of messages and forwarding score of messages separated in control group and treatment group. Data was re-sampled to 1-minute bins aggregating by sum. Number of WhatsApp groups at control was $N = 5.084$ and treatment group had $N = 329$. Were normalized in order to match $N = 100$ for both groups. Therefore *control_q* and *treat_q* represents the number of messages sent in 100 groups each minute. And *control_fs* and *treat_fs* represent the aggregated number of forwarding score of messages sent in 100 groups each minute. The columns show the number of observations, mean, standard deviation, minimum value, .25, .5 and .75 quantils and the maximum value.

4 METHODOLOGY

4.1 RD design in political economy

On Novembre 1st of 2022, the Supreme Electoral Court (TSE) of Brazil issued a judicial decision to WhatsApp ordering the banning of groups engaging with anti-democratic activities[1]. It was possible to verify that WhatsApp took down a cluster of those groups at a specific time, 00:32 of Novemeber 2nd. This rapid decision with no previous knowledge of users in WhatsApp could have produced a quasi-experiment that can be explored to estimate the effects of judicial decision on a sample of anti-democratic groups that were not turned off.

A *regression-discontinuity design* (RDD) approach was chosen to investigate causal relationship between court decision and a change in user behaviour on WhatsApp. Thistlethwaite and Campbell (1960) first applied the use of regression-discontinuity analysis testing causal hypotheses on educational psychology research. Angrist and Lavy (1999) work on Maimmonaide's rule studied the effect of class size on student performance. Since then, RDD use widespread through micro-econometric literature to evaluate average treatment effects (ATE) Cook (2008).

Seminal work of Lee (2008) pioneered using margin of victory as running variable in RDD, establishing the weak conditions in which causal inferences from quasi-experiment using RDD can be as credible as randomized experiment. Looking at mayoral brazilian elections rule, Fujiwara (2011) explored discontinuity in the assignment of single-ballot and dual-ballot to test causal validity of Duverger's Law. Bruce et al. (2021) used discontinuity in close mayoral races between male and female candidates and found that female leadership outperformed males when dealing with health policies in the COVID-19 pandemic.

Cunningham (2021) tells that in order estimate ATE with RD design, is required

[1] https://www1.folha.uol.com.br/poder/2022/11/bolsonaristas-e-tse-fazem-jogo-de-gato-e-rato-com-
shtml

that the continuity assumption is valid. This means that the only change in outcome needs to come from abrupt change at cohort. Lee and Lemieux (2010) and Imbens and Lemieux (2008) provides a best-practice guide to RD design for empirical research, RD can be invalid if individuals can precisely manipulate assignment variable. McCrary (2008) density test helps verifying sorting and anticipation problems with RD. These conditions may be sufficient with cross-sectional framework to capture causal effects with RD, however the discontinuity explored in this work uses time as running variable.

4.2 Potential bias of RDiT and identification strategy

Exploring discontinuity with cohort at a specific time requires identification strategy for "regression-discontinuity in time" (RDiT). Research using RDiT has been made in public economics, environmental economics, industrial organization and many other fields. Anderson (2014) measured effect of subway strikes on traffic congestion. Davis (2008) also uses this design to measure regulation change in driving law in Mexico City. Auffhammer and Kellogg (2011) uses RDiT analyzing the effects of gasoline regulation on air quality.

Hausman and Rapson (2017) give instructions in which conditions RDiT can be valid to estimate causal effects. They highlight three different potential bias using RDiT:

1. **Time-Varying Treatment Effects:** if treatment occur in different periods of time, specification of treatment variable would be incorrect. Since all banned groups were turned off at the same time by WhatsApp, this is not valid in this case. Different time-range windows and frequency bins can be used to increase robustness of analysis;

2. **Autoregression:** if standard errors are correlated with time-dependent variable it could produce bias estimate. Figure 3 suggested that there is an intra-daily cycle. The day of the week could also have an effect on user activity. There are different strategies to address this problem like removing lagged differences or controlling for specific cycles;

3. **Sorting and anticipation effects:** if WhatsApp users could anticipate group shut down, they would be able to reorganize themselves in other groups previous to the turn off. If this happened, then estimate would be biased. This is not a possibility since court decision is unprecedented and it only became public days after the shutdown.

A representation of the causal graph is shown in Figure 5, where T is the treatment (shut down) applied to the banned groups B. Since the sample doesn't contain all WhatsApp groups, B can affect unobserved groups U_t related to the theme, that can therefore affect our selected anti-democratic groups that were not banned Y_t. There are also unobserved groups U_c that has no political discussion, these groups may impact behaviour in selected control group Y_c and Y_t, since these miscellaneous type of messages can appear on all type of groups. Important to notice that treatment B can impact Y_t directly or through U_t. It is assumed that U_t and B have no effect on Y_c because there were no tagged anti-democratic content found in those groups.

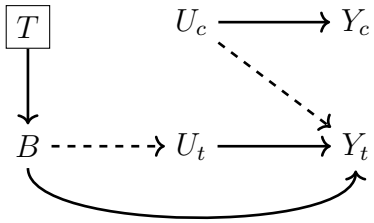

Figure 5: DAG (Directed acyclic graph) representation of causal model proposed. T represents treatment imposed to B (banned anti-democratic groups). U_t are unobserved anti-democratic WhatsApp groups that were not turned off. U_c are unobserved WhatsApp groups that have no relation to anti-democratic mobilization.Y_t are observed anti-democratic WhatsApp groups and Y_c are observed common groups that had no relation to anti-democratic mobilization.

To measure the effects produced by T, RDiT will be used with date as a running variable and 00:32 a.m. of November 2nd as the discontinuity threshold. Following Imbens and Lemieux (2008), the model was specified as:

$$y_{it} = \alpha + \beta \cdot threshold_{it} + f(date_{it}) + \varepsilon_{it} \tag{4.1}$$

In (1), y_{it} is the average change in activity (quantity or forwarding score) for group i (treatment or control) during time t, $threshold_{it}$ is the threshold variable, and $date_{it}$ is the difference of selected frequency from the discontinuity. Model can be estimated with $date_{it}$ being set to zero at the moment of the measure:

$$y_{it} = \alpha + \beta \cdot threshold_{it} + \gamma_1 \cdot date_{it} + \gamma_2 \cdot date_{it} \cdot threshold_{it} + \varepsilon_{it} \tag{4.2}$$

Using similar specification from Anderson (2014), the function $f(date_{it})$ from (1) is defined as $\gamma_1 \cdot date_{it} + \gamma_2 \cdot date_{it} \cdot threshold_{it}$ in (2), where the second term should absorb any smooth relationship between the date and ε_{it} in the moments near the shutdown.

Applying this RD design will help extract trough β the impact on WhatsApp of the TSE decision.

5 RESULTS

Autoregression is the only one of the three potential bias of RDiT that could affect this investigation. Hausman and Rapson (2017) recommends using different time-windows, controlling for variable in estimation or differentiating the lagged dependent variable. To separate possible short-run and long-run effects, different time-windows and time-bin frequencies will be selected. To estimate short-run effect I will use and 30-minutes bins and 6-days window with cohort at the center. To measure medium-run effects I will use 2-hour bin with 28-days window. To remove daily cycles, lagged differences will be applied to the same time of the previous day, therefore measuring variation of dependent variable.

5.1 Short-run effects

Table 3 presents estimates for short-run effects of TSE decision execution on WhatsApp groups that weren't turned off. Running variable is $date_{it}$ as described in methodology section. For this estimate, the dependent variable is the average change in quantity of messages sent in 30-minutes bin. Control (1) are groups of various themes and Treatment (2) are anti-democratic groups. The RD estimate *threshold* shows no statistically significant results to any group, which suggests that court decision had no impact on quantity variation on any kind of WhatsApp group.

Figures 6 and 7 are scatter plots with predicted change in quantity based on short-run RDiT estimate from Table 3. Figure 6 with green points represents control group, whilst Figure 7 with blue points represents treatment group. Looking at the predicted curves and the trajectory of bins through time, it's not possible to verify any substantial change in message flow in short-term after execution of the judicial decision.

Table 4 presents estimates for short-run effects of TSE decision execution on WhatsApp groups that weren't turned off. Like in Table 3, running variable is $date_{it}$, but for this estimate, the dependent variable is the average change in forwarding score of messages sent in 30-minutes bin. Control (1) are groups of various themes and Treatment (2)

are anti-democratic groups. The RD estimate *threshold* shows no statistically significant results to the control group (1), which suggests that court decision had no impact on forwarding scores variation on those groups. However, the treatment group (2) estimate presents a decrease of -1068.76 in forwarding score, with high statistical significance. The interpretation for this value is that in a sample of 100 anti-democratic groups, forwarded messages reached at least -1068 people every 30 minutes shortly after the shutdown. It's important to say "at least" because of the way forwarding score is computed by WhatsApp truncates maximum value at 127.

Figures 8 and 9 are scatter plots with predicted change in forwarding score based on short-run RDiT estimate from Table 4. Figure 8 with green points represents control group, whilst Figure 9 with blue points represents treatment group. As noticed from Table 4 estimates, forwarding score curve remained completely flat for the control group (1). Looking at Figure 9, the discontinuity is verified with the reduction of forwarding score. Since there is no change in users from anti-democratic groups activity at the time (Table 3 and Figure 7) and no apparent change in forwarding score from control group (Table 4 and Figure 8), continuity assumption required required by Cunningham (2021) seems to be valid.

Table 3: Short-run effects of groups shutdown on quantity of messages

	Average change in quantity	
	Control (1)	Treatment (2)
date	0.158***	1.304***
	(0.056)	(0.459)
threshold	5.412	-36.022
	(3.306)	(26.997)
threshold:date	-0.275***	-1.403**
	(0.080)	(0.649)
Observations	288	288
R^2	0.092	0.028
Adjusted R^2	0.082	0.018
Residual Std. Error	14.024(df = 284)	114.516(df = 284)
F Statistic	9.551*** (df = 3.0; 284.0)	2.777** (df = 3.0; 284.0)

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

This table presents short-run RDiT estimates of the effect of shutting down anti-democratic WhatsApp groups on the ones that were not turned off. Robust standard errors in parentheses. The dependent variable in columns (1) and (2) is the average change in quantity flow of messages for 100 groups in each 30 minutes. Control (1) groups are the ones with no relation to political debate, and treatment (2) are anti-democratic groups. The running variable is *date*, *threshold* is the RD estimator and *threshold* : *date* is smoothing term. Cohort was set to November 2nd at 00:32;

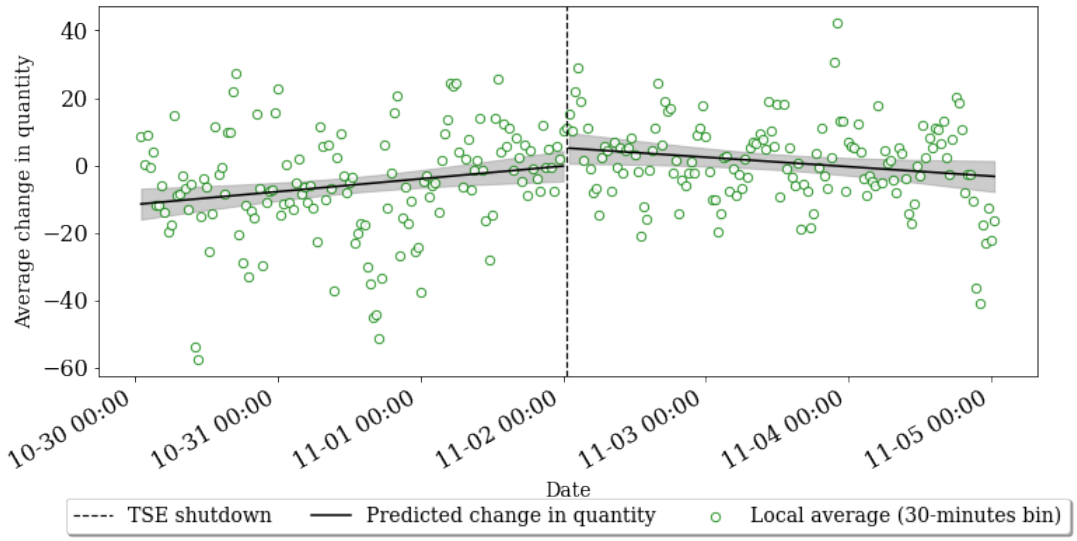

Figure 6: Control group: The short-run effect of court decision on quantity of messages sent on WhatsApp groups of various themes. The vertical axis measures the average change in quantity of messages compared to same time in the previous day. The horizontal axis is the running variable *date* measured in hours from cohort (increases .5 each observation). The dashed vertical line represents the exact time of the execution of TSE decision at 00:32 on November 2nd.

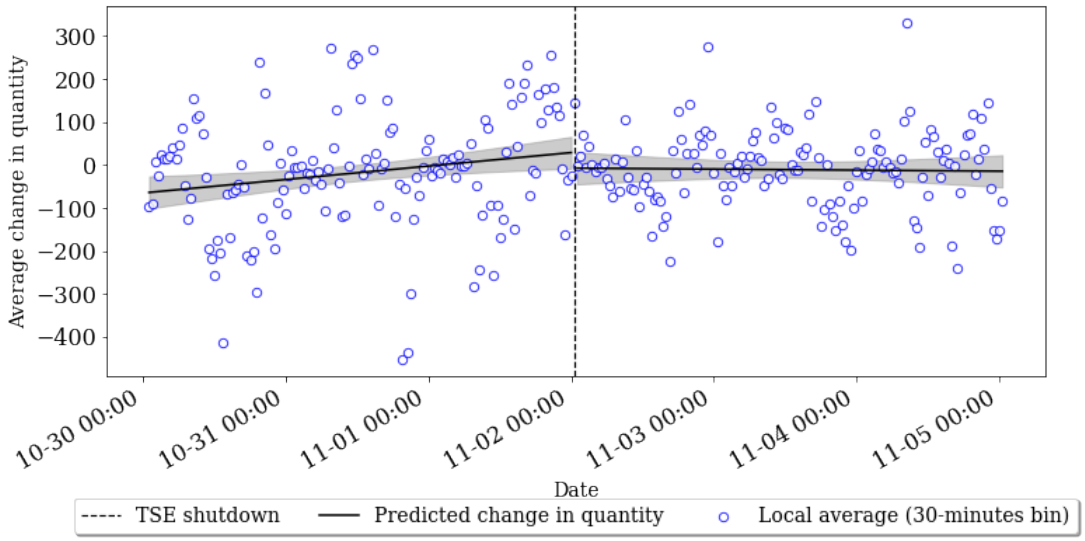

Figure 7: Treatment group: The short-run effect of court decision on quantity of messages sent on anti-democratic WhatsApp groups. The vertical axis measures the average change in quantity of messages compared to same time in the previous day. The horizontal axis is the running variable *date* measured in hours from cohort (increases .5 each observation). The dashed vertical line represents the exact time of the execution of TSE decision at 00:32 on November 2nd.

Table 4: Short-run effects of groups shutdown on forwarding score of messages

	Average change in forwarding score	
	Control (1)	Treatment (2)
date	0.282*	13.951***
	(0.155)	(2.725)
threshold	1.102	-1068.760***
	(9.103)	(160.248)
threshold:date	-0.320	-8.458**
	(0.219)	(3.854)
Observations	288	288
R^2	0.027	0.155
Adjusted R^2	0.017	0.146
Residual Std. Error	38.615(df = 284)	679.751(df = 284)
F Statistic	2.624* (df = 3.0; 284.0)	17.396*** (df = 3.0; 284.0)
Note:		*p<0.1; **p<0.05; ***p<0.01

This table presents short-run RDiT estimates of the effect of shutting down anti-democratic WhatsApp groups on the ones that were not turned off. Robust standard errors in parentheses. The dependent variable in columns (1) and (2) is the average change in forwarding score of messages for 100 groups in each 30 minutes. Control (1) groups are the ones with no relation to political debate, and treatment (2) are anti-democratic groups. The running variable is *date*, *threshold* is the RD estimator and *threshold* : *date* is smoothing term. Cohort was set to November 2nd at 00:32;

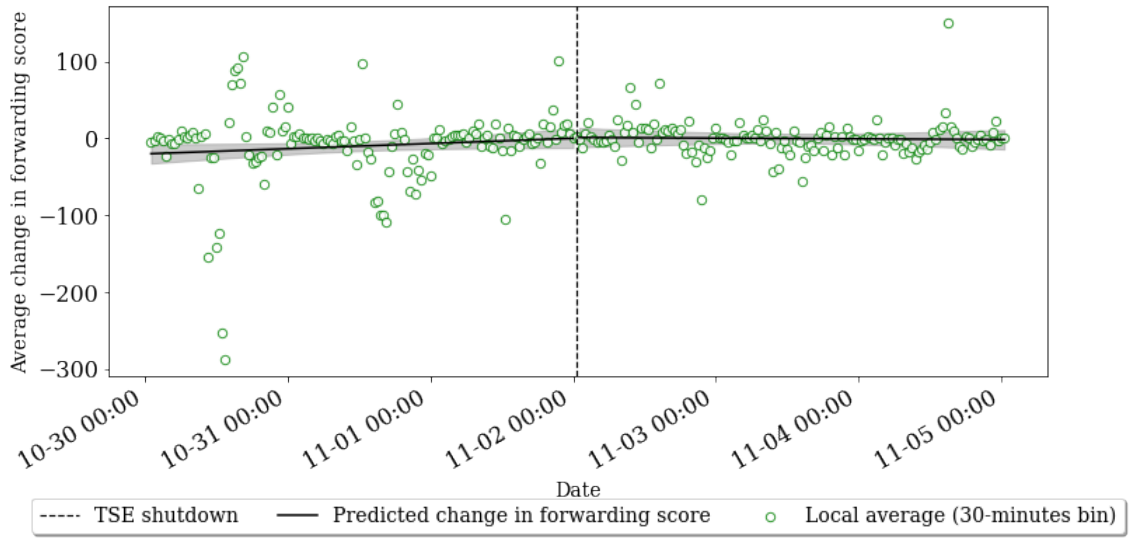

Figure 8: Control group: The short-run effect of court decision on forwarding score of messages sent on WhatsApp groups of various themes. The vertical axis measures the average change in forwarding score of messages compared to same time in the previous day. The horizontal axis is the running variable *date* measured in hours from cohort (increases .5 in each observation). The dashed vertical line represents the exact time of the execution of TSE decision at 00:32 on November 2nd.

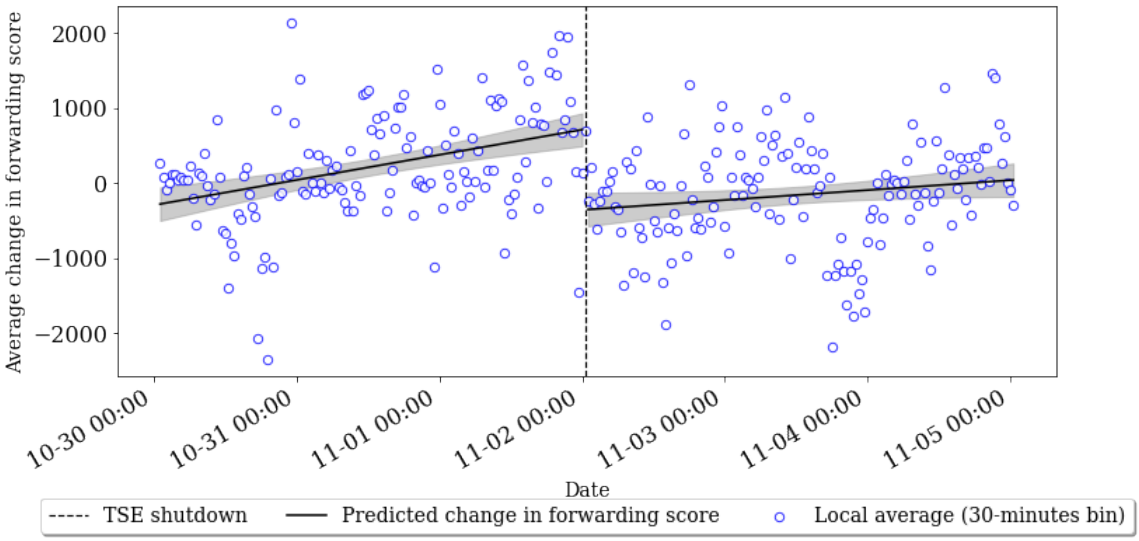

Figure 9: Treatment group: The short-run effect of court decision on forwarding score of messages sent on anti-democratic WhatsApp groups. The vertical axis measures the average change in forwarding score of messages compared to same time in the previous day. The horizontal axis is the running variable *date* measured in hours from cohort (increases .5 in each observation). The dashed vertical line represents the exact time of the execution of TSE decision at 00:32 on November 2nd.

5.2 Medium-run effects

A longer window of time with wider frequency-bin was selected to make more robust conclusions on possible effects of deplatforming measure. Table 5 presents estimates for medium-run effects of TSE decision execution on the same groups from Tables 3 and 4. Running variable is still $date_{it}$, and the dependent variable is the average change in quantity of messages sent in 2-hours bin. Such as before, Control (1) are groups of various themes and Treatment (2) are anti-democratic groups. The RD estimate *threshold* shows no statistically significant results that impact dependent variables, which suggests that court decision had no impact on quantity variation on any of those groups.

Figures 10 and 11 are scatter plots with predicted change in quantity based on medium-run RDiT estimate from Table 5. Like in the previous figures, Figure 10 with green points represents control group and Figure 11 with blue points represents treatment group. As noticed from Table 5 estimates, quantity of messages had no change in any of the two plots

Table 6 presents the results of RD estimate for medium-run effects with forwarding score as dependent variable. Control group (1) shows no change at *threshold*, remaining stable. But again, at the treatment group (2), estimate presents a statistically significant decrease of -1589.929 in forwarding score. This value suggests that in a sample of 100 anti-democratic groups, forwarded messages reached at least -1589 people every 2 hours shortly after the shutdown. This result confirms discontinuity found in Table 4 with 30-minutes frequency.

Figures 12 and 13 follow the same standard of the other ones. They present predicted change in forwarding score based on medium-run RDiT estimate from Table 6. Figure 12 with green points represents control group and Figure 13 with blue points represents treatment group. Figure 12 shows no discontinuity, with predicted value remaining stable. Figure 13 show positive trend of forwarding score with an abrupt reduction at cuttoff. Like in the short-run case, continuity assumption seems to be valid for medium-run estimate. This result suggest that TSE decision had an effect reducing the reach of forwarded messages for users in anti-democratic.

Table 5: Medium-run effects of groups shutdown on quantity of messages

	Average change in quantity	
	Control (1)	Treatment (2)
date	0.071	-0.151
	(0.067)	(0.320)
threshold	1.171	-74.650
	(15.848)	(75.326)
threshold:date	-0.172*	0.726
	(0.095)	(0.453)
Observations	288	288
R^2	0.012	0.012
Adjusted R^2	0.002	0.002
Residual Std. Error	67.233(df = 284)	319.569(df = 284)
F Statistic	1.189 (df = 3.0; 284.0)	1.186 (df = 3.0; 284.0)

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

This table presents medium-run RDiT estimates of the effect of shutting down anti-democratic WhatsApp groups on the ones that were not turned off. Robust standard errors in parentheses. The dependent variable in columns (1) and (2) is the average change in quantity of messages for 100 groups in each 2 hours. Control (1) groups are the ones with no relation to political debate, and treatment (2) are anti-democratic groups. The running variable is *date*, *threshold* is the RD estimator and *threshold* : *date* is smoothing term. Cohort was set to November 2nd at 00:32;

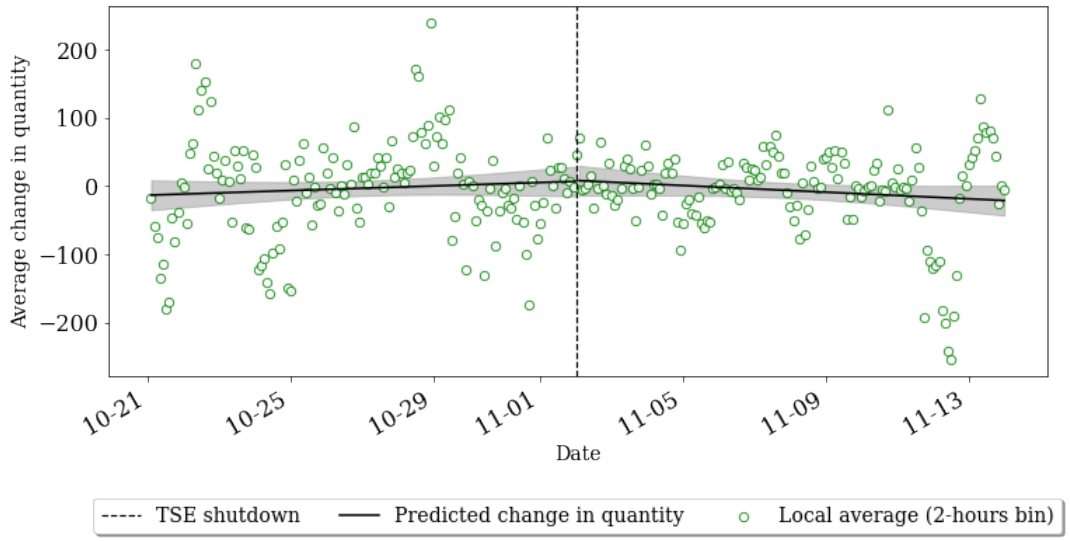

Figure 10: Control group: The medium-run effect of court decision on quantity of messages sent on WhatsApp groups of various themes. The vertical axis measures the average change in quantity of messages compared to same time in the previous day. The horizontal axis is the running variable *date* measured in hours from cohort (increases 2 in each observation). The dashed vertical line represents the exact time of the execution of TSE decision at 00:32 on November 2nd.

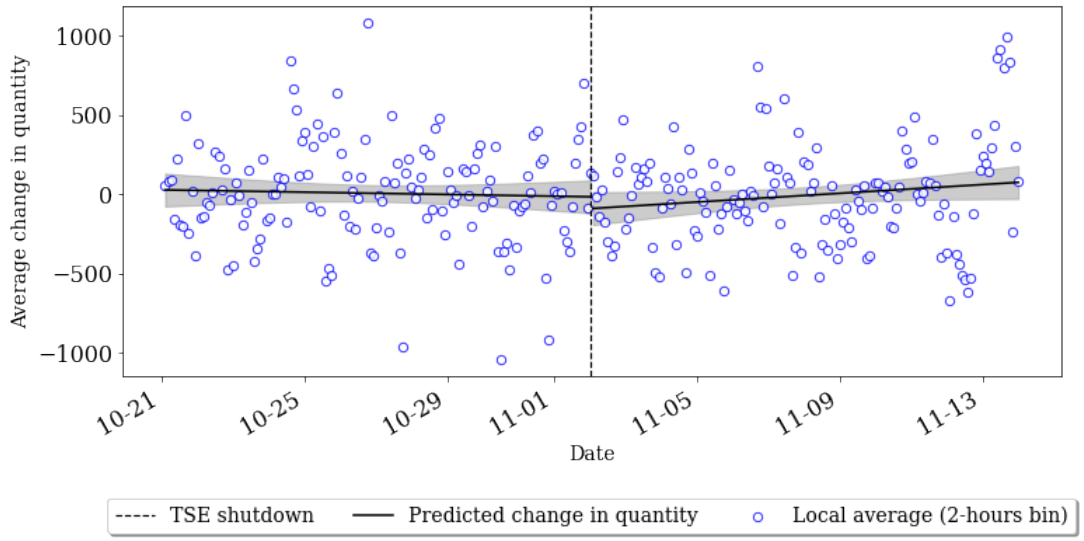

Figure 11: Treatment group: The medium-run effect of court decision on quantity of messages sent on anti-democratic WhatsApp groups. The vertical axis measures the average change in quantity of messages compared to same time in the previous day. The horizontal axis is the running variable *date* measured in hours from cohort (increases 2 in each observation). The dashed vertical line represents the exact time of the execution of TSE decision at 00:32 on November 2nd.

Table 6: Medium-run effects of groups shutdown on forwarding score of messages

	Average change in forwarding score	
	Control (1)	Treatment (2)
date	-0.028	2.829
	(0.124)	(1.969)
threshold	7.344	-1589.929***
	(29.081)	(462.985)
threshold:date	0.042	2.534
	(0.175)	(2.784)
Observations	288	288
R^2	0.001	0.042
Adjusted R^2	-0.010	0.032
Residual Std. Error	123.375(df = 284)	1964.209(df = 284)
F Statistic	0.065 (df = 3.0; 284.0)	4.196*** (df = 3.0; 284.0)

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

This table presents medium-run RDiT estimates of the effect of shutting down anti-democratic WhatsApp groups on the ones that were not turned off. Robust standard errors in parentheses. The dependent variable in columns (1) and (2) is the average change in forwarding score of messages for 100 groups in each 2 hours. Control (1) groups are the ones with no relation to political debate, and treatment (2) are anti-democratic groups. The running variable is *date*, *threshold* is the RD estimator and *threshold* : *date* is smoothing term. Cohort was set to November 2nd at 00:32;

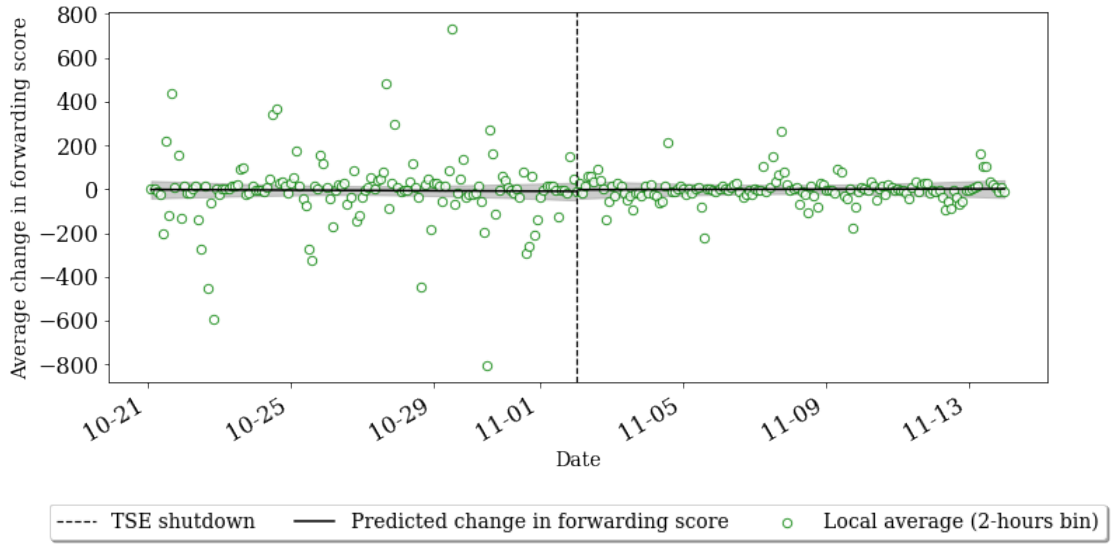

Figure 12: Control group: The medium-run effect of court decision on forwarding score of messages sent on WhatsApp groups of various themes. The vertical axis measures the average change in forwarding score of messages compared to same time in the previous day. The horizontal axis is the running variable *date* measured in hours from cohort (increases 2 in each observation). The dashed vertical line represents the exact time of the execution of TSE decision at 00:32 on November 2nd.

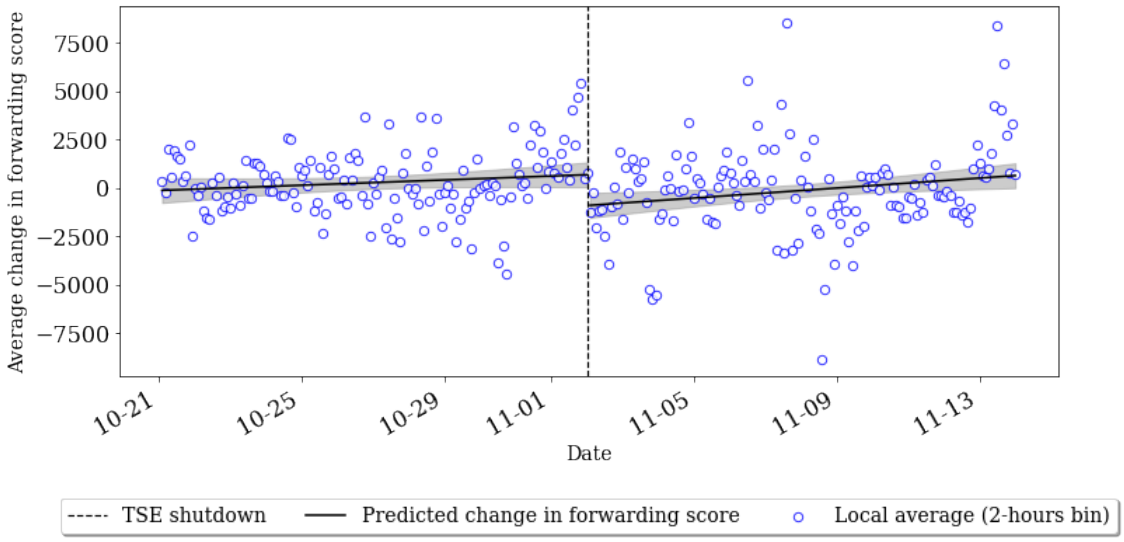

Figure 13: Treatment group: The medium-run effect of court decision on forwarding score of messages sent on anti-democratic WhatsApp groups. The vertical axis measures the average change in forwarding score of messages compared to same time in the previous day. The horizontal axis is the running variable *date* measured in hours from cohort (increases 2 in each observation). The dashed vertical line represents the exact time of the execution of TSE decision at 00:32 on November 2nd.

5.3 Discussion

The results from estimates with quantity of messages as dependent variable addresses one of the objectives of this dissertation. Weather the Supreme Electoral Court (TSE) decision could have suppressed users from using WhatsApp to converse. Rogers (2020) prompted the possibility that deplatforming increased anti-establishment sentiment. Preventing people from communicating can be seen as an attempt to silence free-speech. The results presented in Table 3 and 5 measure the average change in quantity of messages sent. No reduction in user activity was found in any type of WhatsApp groups. It is possible to say that conversation wasn't affected in those groups, quantity of messages showed no significant change after threshold.

When looking to forwarded messages, the results are a bit different. While no change at threshold is seen on control group, a discontinuity lowering activity was found for treatment group in both shot-run and medium-run estimates. A reduction in forwarding score could mean that the channels in which viral messages travel through were affected in someway. This result addressed the two other specific objectives of this work. This reduction was significant on *threshold* and could represent at least minus 1068 people reached for each 30-minutes on short-run and minus 1590 every 2-hours for medium-run.

There was no change in user activity in all groups monitored that could explain a reduction of forwarded messages. Reduction in forwarding score must have been caused by factors external to those groups. The reduction verified only occurred for groups that had shared anti-democratic content and engaged with extremist narrative. When looking at Figures 9 and 13 it seems that the trajectory of forwarding messages was ascending. All these facts combined, leads to believe that reduction of forwarding score was caused by the groups shutdown.

This work is a simplified first attempt to understand how regulations can impact WhatsApp. Further investigation could look in to how better classification of ideological profile of the users can improve analysis. Exploring this same discontinuity, it's also possible to look if there are any regional effects or even if the type of viral message react differently depending on the media type (images, audios, videos), duration, length of text or other types of data available to explore. Another step would be evaluating if this decision produced spillovers to other social media like Twitter or Facebook and how users reorganized themselves.

6 CONCLUSION

WhatsApp is the most used social media in Brazil and it had an important role in 2018 and 2022 elections. Days after the election result confirmation on October 30, anti-democratic riots began appearing in front of military quarters asking for an intervention on results. For the first time, the Supreme Electoral Court (TSE) issued a decision making WhatsAppp shut down the cluster of groups promoting anti-democratic riots. This work used data from public WhatsApp groups to measure the impact of this decision on anti-democratic groups that were not banned.

Results suggests that the shut down of the anti-democratic groups had no effect on the quantity of messages exchanged in any type of group, but when looking to forwarded messages, anti-democratic groups presented a reduction. Viral messages could have some type of inertial movement when travelling through groups. When these paths were closed at the same time, velocity and reach of the viral messages was reduced in these groups. Understanding the effects of possible mechanism to counter extremism is important to help policy makers design better institutional resources to regulate social media.

By the time I am finishing this work, Donald Trump and many other far-right "influencers" were rehabilitated on Twitter[1]. Brazilian judicial system is still struggling with anti-democratic organization on social media[2]. These events shows the urgency on developing effective tools for government institutions to counter extremism without damaging individual freedoms and the functioning technology.

[1] https://oglobo.globo.com/economia/tecnologia/noticia/2022/11/apos-trump-twitter-reabilita-ye-perfis-polemicos-podem-afastar-ainda-mais-os-anunciantes.ghtml

[2] https://oglobo.globo.com/blogs/sonar-a-escuta-das-redes/post/2022/11/nos-grupos-de-whatsapp-bolsonaristas-miram-moraes-e-esquecem-lula-mostra-monitoramento.ghtml

BIBLIOGRAPHY

Helen Margetts. 421The Internet and Democracy. In *The Oxford Handbook of Internet Studies*. Oxford University Press, 01 2013. ISBN 9780199589074. doi: 10.1093/oxfordhb/9780199589074.013.0020. URL https://doi.org/10.1093/oxfordhb/9780199589074.013.0020.

Michael L. Best and Keegan W. Wade. The internet and democracy: Global catalyst or democratic dud? *Bulletin of Science, Technology & Society*, 29(4):255–271, 2009. doi: 10.1177/0270467609336304. URL https://doi.org/10.1177/0270467609336304.

Jim Isaak and Mina J. Hanna. User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer*, 51(8):56–59, 2018. doi: 10.1109/MC.2018.3191268.

Adam Przeworski, Susan C. Stokes, and Bernard Manin, editors. *Democracy, Accountability, and Representation*. Cambridge University Press, September 1999. doi: 10.1017/cbo9781139175104. URL https://doi.org/10.1017/cbo9781139175104.

Elliott Ash, Ruben Durante, Maria Grebenshchikova, and Carlo Schwarz. Visual stereotypes in news media. *SSRN Electronic Journal*, 2021. doi: 10.2139/ssrn.3934858. URL https://doi.org/10.2139/ssrn.3934858.

Alexsandros Cavgias, Raphael Corbi, Luis Meloni, and Lucas M. Novaes. EDITED DEMOCRACY: Media Manipulation and the News Coverage of Presidential Debates. Technical report, May 2019. URL https://ideas.repec.org/p/spa/wpaper/2019wpecon17.html.

Gregory J. Martin and Ali Yurukoglu. Bias in cable news: Persuasion and polarization. *American Economic Review*, 107(9):2565–2599, September 2017. doi: 10.1257/aer.20160812. URL https://doi.org/10.1257/aer.20160812.

Oliver Falck, Robert Gold, and Stephan Heblich. E-lections: Voting behavior and the internet. *American Economic Review*, 104(7):2238–2265, July 2014. doi: 10.1257/aer.104.7.2238. URL https://doi.org/10.1257/aer.104.7.2238.

Alessandro Bessi, Fabiana Zollo, Michela Del Vicario, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Trend of narratives in the age of misinformation. *PLOS*

ONE, 10(8):e0134641, August 2015. doi: 10.1371/journal.pone.0134641. URL https://doi.org/10.1371/journal.pone.0134641.

Karsten Müller and Carlo Schwarz. Fanning the flames of hate: Social media and hate crime. *SSRN Electronic Journal*, 2017. doi: 10.2139/ssrn.3082972. URL https://doi.org/10.2139/ssrn.3082972.

Karsten Müller and Carlo Schwarz. Making america hate again? twitter and hate crime under trump. *SSRN Electronic Journal*, 2018. doi: 10.2139/ssrn.3149103. URL https://doi.org/10.2139/ssrn.3149103.

Thomas Fujiwara, Karsten Müller, and Carlo Schwarz. The effect of social media on elections: Evidence from the united states. Technical report, May 2021. URL https://doi.org/10.3386/w28849.

Kokil Jaidka, Alvin Zhou, and Yphtach Lelkes. Brevity is the soul of twitter: The constraint affordance and political discussion. *Journal of Communication*, 69(4):345–372, July 2019. doi: 10.1093/joc/jqz023. URL https://doi.org/10.1093/joc/jqz023.

Ro'ee Levy. Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, 111(3):831–870, March 2021. doi: 10.1257/aer.20191777. URL https://doi.org/10.1257/aer.20191777.

Sarah T. Roberts. *Behind the Screen*. Yale University Press, June 2019. doi: 10.2307/j.ctvhrcz0v. URL https://doi.org/10.2307/j.ctvhrcz0v.

Tamar Mitts. Banned: How deplatforming extremists mobilizes hate in the dark corners of the internet. 2021.

Adrian Rauchfleisch and Jonas Kaiser. Deplatforming the far-right: An analysis of YouTube and BitChute. *SSRN Electronic Journal*, 2021. doi: 10.2139/ssrn.3867818. URL https://doi.org/10.2139/ssrn.3867818.

Gabriel Luis Santos Freire, Tales Panoutsos, Lucas Perez, Fabricio Benevenuto, and Flavio Figueiredo. Understanding effects of moderation and migration on online video sharing platforms. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*. ACM, June 2022. doi: 10.1145/3511095.3536377. URL https://doi.org/10.1145/3511095.3536377.

Manoel Horta Ribeiro, Justin Cheng, and Robert West. Automated content moderation increases adherence to community guidelines, 2022. URL https://arxiv.org/abs/2210.10454.

Richard Rogers. Deplatforming: Following extreme internet celebrities to telegram and alternative social media. *European Journal of Communication*, 35(3):213–229, May 2020. doi: 10.1177/0267323120922066. URL https://doi.org/10.1177/0267323120922066.

Thomas Fujiwara. Voting technology, political responsiveness, and infant health: Evidence from brazil. *Econometrica*, 83(2):423–464, 2015. doi: 10.3982/ecta11520. URL https://doi.org/10.3982/ecta11520.

Rafael Evangelista and Fernanda Bruno. WhatsApp and political instability in brazil: targeted messages and political radicalisation. *Internet Policy Review*, 8(4), December 2019. doi: 10.14763/2019.4.1434. URL https://doi.org/10.14763/2019.4.1434.

Philipe Melo, Fabrício Benevenuto, Daniel Kansaon, Vitor Mafra, and Kaio Sá. Monitor de WhatsApp: Um sistema para checagem de fatos no combate à desinformação. In *Anais Estendidos do XXVII Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia 2021)*. Sociedade Brasileira de Computação - SBC, November 2021. doi: 10.5753/webmedia_estendido.2021.17617. URL https://doi.org/10.5753/webmedia_estendido.2021.17617.

Gustavo Resende, Johnnatan Messias, Márcio Silva, Jussara Almeida, Marisa Vasconcelos, and Fabrício Benevenuto. A system for monitoring public political groups in WhatsApp. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*. ACM, October 2018. doi: 10.1145/3243082.3264662. URL https://doi.org/10.1145/3243082.3264662.

Matthias Hartung, Roman Klinger, Franziska Schmidtke, and Lars Vogel. Identifying right-wing extremism in german twitter profiles: A classification approach. In *Natural Language Processing and Information Systems*, pages 320–325. Springer International Publishing, 2017. doi: 10.1007/978-3-319-59569-6_40. URL https://doi.org/10.1007/978-3-319-59569-6_40.

Eoin Lenihan. A classification of antifa twitter accounts based on social network mapping and linguistic analysis. *Social Network Analysis and Mining*, 12(1), November 2021. doi: 10.1007/s13278-021-00847-8. URL https://doi.org/10.1007/s13278-021-00847-8.

Caio Machado, Beatriz Kira, Vidya Narayanan, Bence Kollanyi, and Philip Howard. A study of misinformation in WhatsApp groups with a focus on the brazilian presidential elections. In *Companion Proceedings of The 2019 World Wide Web Conference*. ACM,

May 2019. doi: 10.1145/3308560.3316738. URL https://doi.org/10.1145/3308560.3316738.

Donald L. Thistlethwaite and Donald T. Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6): 309–317, December 1960. doi: 10.1037/h0044319. URL https://doi.org/10.1037/h0044319.

J. D. Angrist and V. Lavy. Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, 114(2):533–575, May 1999. doi: 10.1162/003355399556061. URL https://doi.org/10.1162/003355399556061.

Thomas D. Cook. "waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, 142(2):636–654, February 2008. doi: 10.1016/j.jeconom.2007.05.002. URL https://doi.org/10.1016/j.jeconom.2007.05.002.

David S. Lee. Randomized experiments from non-random selection in u.s. house elections. *Journal of Econometrics*, 142(2):675–697, February 2008. doi: 10.1016/j.jeconom.2007.05.004. URL https://doi.org/10.1016/j.jeconom.2007.05.004.

Thomas Fujiwara. A regression discontinuity test of strategic voting and duverger's law. *Quarterly Journal of Political Science*, 6(3-4):197–233, December 2011. doi: 10.1561/100.00010037. URL https://doi.org/10.1561/100.00010037.

Raphael Bruce, Alexsandros Cavgias, Luis Meloni, and Mario Remigio. Under Pressure: Women's Leadership During the COVID-19 Crisis. Working Papers, Department of Economics $2021_19, University of São Paulo (FEA-USP), July 2021. URL$.

Scott Cunningham. *Causal Inference: The Mixtape*. Yale University Press, 2021. ISBN 9780300251685. URL http://www.jstor.org/stable/j.ctv1c29t27.

David S Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355, June 2010. 10.1257/jel.48.2.281. URL https://doi.org/10.1257/jel.48.2.281.

Guido W. Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635, February 2008. 10.1016/j.jeconom.2007.05.001. URL https://doi.org/10.1016/j.jeconom.2007.05.001.

Justin McCrary. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714, February 2008. 10.1016/j.jeconom.2007.05.005. URL https://doi.org/10.1016/j.jeconom.2007.05.005.

Michael L. Anderson. Subways, strikes, and slowdowns: The impacts of public transit on traffic congestion. *American Economic Review*, 104(9):2763–2796, September 2014. 10.1257/aer.104.9.2763. URL https://doi.org/10.1257/aer.104.9.2763.

Lucas W. Davis. The effect of driving restrictions on air quality in mexico city. *Journal of Political Economy*, 116(1):38–81, February 2008. 10.1086/529398. URL https://doi.org/10.1086/529398.

Maximilian Auffhammer and Ryan Kellogg. Clearing the air? the effects of gasoline content regulation on air quality. *American Economic Review*, 101(6):2687–2722, October 2011. 10.1257/aer.101.6.2687. URL https://doi.org/10.1257/aer.101.6.2687.

Catherine Hausman and David Rapson. Regression discontinuity in time: Considerations for empirical applications. Technical report, July 2017. URL https://doi.org/10.3386/w23602.