Essays on Political Economy, Social Norms, and Networks

by

Alejandra Agustina Martínez Martínez

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Economics

Universidad Carlos III de Madrid

Advisor:

Antonio Cabrales Goitia

July 2023

A quienes me dieron todo, mi mamá y mi abuela.

Acknowledgements

First and foremost, I thank Antonio Cabrales for his advice throughout all these years. He was always available and ready to help and support me whenever needed. He actively guided me to discover and develop my research style, respecting my freedom, trusting in me and my work, and putting my thesis within limits when required. Antonio went beyond the limits of an academic advisor-advisee relationship, always recognizing me as a human being with feelings and needs beyond the Ph.D., and took care of me in all dimensions.

I am forever grateful to Johannes Schneider and Warn Lekfuangfu for giving me the strongest support and sharpest advice, especially during this last year of my Ph.D. Both of you were always willing to get into thought-provoking conversations and challenge me to improve all my academic skills, taking care of all the small details that make a paper grow. Thank you for being critical and picky with my work and, at the same time, being supportive and lifting me up when I needed it. I could not imagine a better team and role model(s) than the three of you; Antonio, Johannes, and Warn.

I would not be here without Esteban Nicolini, my first mentor, and dearest friend. I have had many coffees during my life, but only one changed it – the coffee you invited me to talk about the future and asked: *Agustina, what do you think about doing a Ph.D.?* I am grateful for these six years of coffee and talks in Madrid, that fed both my mind and soul.

I would like to thank all faculty members of the Department of Economics at Universidad Carlos III de Madrid. In particular, I want to thank all participants of the Microeconomics and Applied Economics Reading Groups for creating a relaxed and highly stimulating environment to discuss research and receive feedback. My economic thinking greatly improved from regularly participating in these reading groups. I have especially benefited from the comments and feedback of Luigi Minale, Jan Stuhler, Bruno Pessoa, Jesús Fernández, Daniel Rees, Jesús Carro, and Julio Cáceres (from the applied faculty); and Boris Ginzburg, Antoine Loeper, Nicolas Motz, and Martin Dumav (from the micro faculty). I want to extend my gratitude to Antoine Loeper, Jesús Carro, Johannes Schneider, and Julio Cáceres for their help and guidance in developing my teaching skills, which I consider crucial for the academic work. A big thanks to Angélica for her help and patience in the

administrative tasks during these years, and to Luisa, in her roles of master's director and academic godmother, for all her support during that period.

A big piece of this thesis studies peer effects; I am convinced this is not random. My uc3m peers positively affected (with many stars) my academic performance and individual utility - simply called happiness. Thanks to Álvaro, Fernando, Florencia, Henry, Isabel, Lidia, Morteza, and Telmo, for exacerbating the Ph.D. good moments and alleviating the bad ones. I will be forever grateful to this world for meeting Henry and Florencia, who became my dearest friends. These years were plenty of changes, but their company was almost the unique constant in my life, from period 0 - when we were tourists in Madrid - to T - while I write these lines. Thanks to my dear *bruxas*, Lidia, and Florencia, for making our office my second home. It was a safe place to share our lives while working, show our true selves, and find empathy for one another. Álvaro, I could not make it without your company, support, and admirable and contagious optimism during the program's first years – you played a key role in my assimilation to Spain, and I will always be grateful for it.

Besides, I would like to extend my gratitude to my master peers, especially Iván, Belén, Diego, and Beka, for bringing energy and happiness whenever I talk or see you again. Thanks to the uc3m students from other cohorts I connected with during these years, I learned from all of you. I truly enjoyed the company of Onursal, Javier, Rubén, Michelangelo, and Weifeng (from the older cohorts), and Mariale, Antonio, Joël, Lorenzo, Álvaro, Lukas, Óscar, Facundo, Camila, Federico, Sofía, and Serkan (from the younger cohorts) along the program. I am grateful for getting to know Diego, Laura, and David when they visited uc3m. I have enjoyed our talks and your company, inside and outside the department. A special thank goes to Diego, my outstanding coauthor, whose contributions greatly enhanced this dissertation.

I did not know this when I started, but doing a Ph.D. requires skills that are way more important than purely academic ones. Self-trust, patience, not sabotaging myself, and asking for help (as strong people do) have been the most difficult subjects to master; I could not do it without Aidín's guidance. Taking care of my mental health was the best investment I ever made.

I would not be here without the support of my family and friends; they always give me a reason to feel fortunate in life, indenpendently of how my workday is going. A mi mamá, Norma, por darme alas y enseñarme a volar, incluso sabiendo que esas alas podrían llevarme lejos de nuestro Tucumán natal. Gracias por educarme en un contexto de libertad y apoyo incondicional, que me dieron raíces fuertes y la seguridad de que siempre tendré un hogar repleto de amor al cual volver. A mi abuela, Chichí, por su ternura y empatía que no conocen de límites – siempre encontré en tu sonrisa y tus abrazos la fuente de felicidad más simple y pura. A mis tíos, Antonio y Daniel, a mis primos, Nicolás y Jimena, y a mi madrina, Elina, por su cariño, confianza y apoyo constante. *It takes a village to raise a child*, y mi village está compuesta por personas que quiero, admiro y extraño siempre - todo lo soy, se los debo a ustedes. Gracias a mis tías, Susana y Juani, por acogerme y ser mi familia en España.

Tengo una banda amiga que me aguanta el corazón y desde hace seis años, lo hace a 10.000km de distancia: gracias a Verónica, Daniela y Mariano; Marcelo, Vera, Emiliano, Gerardo y Andrés; Maga, Amalita, Nuria y Victoria; Maru, Silvina, Maxi, Aimé, Mauro y Laura por tantas experiencias compartidas, y por una vida de crecimiento y apoyo mutuo. Especialmente, gracias a Vero, Maru, Maga y Nuria, por visitarme en Madrid y conectar mis dos mundos. Gracias a mis compañeros de salsa, especialmente a Javi, Virginia, Jesús, Cristina y Belén, por ser mi cable a tierra en Madrid, compañeros de tantos planes y cómplices de mi acogida madrileña. Gracias a Santiago, por quererme mucho y muy bien; a tu lado mi vida es más fácil e infinitamente más feliz. Sos la excepción a todas mis reglas estúpidas pero la confirmación a todas las importantes, esencial para construir una relación tan sana y bonita como la nuestra. I am looking forward to our future together.

Published and submitted content

Chapters 1 and 3 of this dissertation, titled *Raise your Voice! Activism and Peer Effects in Online Social Networks* and *Hate Speech and Social Media: Evidence from Bolsonaro's Election in Brazil*, have been previously shared as working papers on my academic webpage with the following links:

https://alejandraagustinamartinez.github.io/files/raise_your_voice.pdf

https://alejandraagustinamartinez.github.io/files/hate_speech_brazil.pdf

Abstract

This dissertation is centered on understanding how social interactions affect social norms and individual behavior, especially in the political and public spheres. This thesis comprises three chapters, in which I combine theoretical and empirical analysis. My approach consists of analyzing the local mechanism driving individual actions, potentially resulting in a global outcome.

Chapter 1 of the dissertation studies theoretically and empirically how peers affect individuals' involvement in political activism through social media platforms. Chapter 2 of the dissertation is a theoretical model analyzing how individuals choose their behavior when confronted with local, endogenous, and discrete social norms. Chapter 3 of the dissertation is an empirical project investigating how information shocks trigger a social norm update and subsequent behavior changes.

Chapter 1. Raise your Voice! Activism and Peer Effects in Online Social Networks. Do peers influence individuals' involvement in political activism? To provide a quantitative answer, I study Argentina's abortion rights debate through Twitter, the social media platform. Pro-choice and pro-life activists coexisted online, and the evidence suggests peer groups were not too polarized. I develop a model of strategic interactions in a network allowing for heterogeneous peer effects. Next, I estimate peer effects and test whether online activism exhibits strategic substitutability or complementarity. I create a novel panel dataset where links and actions are observable by combining tweets' and users' information. I provide a reduced-form analysis by proposing a network-based instrumental variable. The results indicate strategic complementarity in online activism from both aligned and opposing peers. Notably, the evidence suggests homophily in the formation of Twitter's network, but it does not support the hypothesis of an echo-chamber effect.

Chapter 2. Discrete Social Norms in Networks. In this paper, I present a model of social norms and conformity, assuming that individuals are embedded in a social network. Each individual chooses an action, or a code of conduct, based on her preferences and social norm. In the model, social norms are *local and endogenous*, defined as the average behavior of peer groups, and *discrete* - to shedding light on real-world examples, such as female labor force participation or female genital

cutting. I show that when actions are discrete, different social norms emerge, as the game generally admits multiple equilibria. In these cases, despite the local nature of the game, global social norms may arise due to a purely conformist equilibrium in which all individuals choose the same code of conduct. Then, I provide evidence that when a unique equilibrium exists, it displays a high degree of individualism - although certain individuals conform to the norm for some network structures. Precisely, the equilibrium outcome is pinned down by four factors: the distribution of preferences, the network structure, the society's taste for conformity, and the choice set.

Chapter 3. Hate Speech and Social Media: Evidence from Bolsonaro's Election in Brazil. How does newly available information affect individuals' perception of social norms and, consequently, behavior? We examine the impact of Bolsonaro's victory in the 2018 Brazilian presidential election on the prevalence of online hate speech. This project relies on Twitter data from 2017 to 2019 and employs text analysis techniques to detect hate speech in tweets' content. To causally identify the impact of Bolsonaro's election on hate speech through Twitter, we follow a difference-in-differences approach, using the election result as an information shock. We estimate two difference-in-differences models, the traditional and another with a continuous treatment variable. In the latter, the election result at each municipality measures the local incidence of this information shock. Our findings reveal that online hate speech experienced a surge following the elections, particularly in municipalities where Bolsonaro's popularity was relatively low. These results are further supported by individual-level regressions, which show that both extensive and intensive margins of individual hate speech contributed to the overall increase. We interpret these findings within the framework of a belief updating mechanism, specifically emphasizing the process of revising social norms that govern what is acceptable to say (or not) in public.

Contents

List of Figures

List of Tables

Chapter 1

Raise your Voice! Activism and Peer Effects in Online Social Networks

1.1 Introduction

What is the influence of peers on individuals' engagement in *political activism*?[1] There is no straightforward answer to this question - related to a collective action problem. First, there is no theoretical agreement on the strategic nature of activism. Model assumptions on the utility function and information structure determine whether actions are strategic substitutes or complements[2] - Olson (2009), Ostrom (2000), Edmond (2013), Passarelli and Tabellini (2017). Second, empirical research of peer influence on political activism is scarce; some exceptions are Bursztyn et al. (2021), Cantoni et al. (2019), González (2020), and Hager et al. (2023). This scarcity is explained twofold: identifying the influence of peers in individual actions is complex (Manski, 1993) and estimating it requires specific data - including at least a rough approximation of social interactions. In a novel context, this paper contributes to the literature on collective action problems by examining *peer effects*[3] in political activism.

In this paper, I rely on data from Twitter, which provides an ideal context for studying peer effects in political activism. Social media platforms have created a new public sphere where individuals connect, interact, and communicate. As for Twitter, *hashtags* have become a default method to designate online collective thoughts, ideas, and claims. Among them are the ones advocating for social change - constituting the *online version of political activism*: #BlackLivesMatter, #MeToo,

[1] I refer to political activism as the participation in a collective claim demanding political rights.

[2] Scholars usually frame collective action problems as a public good or a coordination game, leading to different implications about the strategic nature of actions.

[3] That is, the influence of peers' actions on individuals' actions.

#LoveIsLove, #ClimateAction. Moreover, Twitter offers precise observability of online links and rich data on social interactions. Regarding the decision to follow an account, unilateral and bilateral ties exist. Users interact in several ways: by posting, replying, retweeting, and quoting tweets.

My approach to investigating how peers affect political activism focuses on understanding the *local and direct mechanism* - the influence of peers' actions - that drives *individual political behavior* and leads to a *global outcome* - a collective claim. To frame this question, I develop a theoretical model of peer effects in a network that explicitly assumes individuals care about their peers' activism. Then, I estimate the model by proposing a network-based instrumental variable, relying on Twitter data to conduct the empirical analysis. I focus on the *intensive margin* of political activism, offering a quantitative measure of activism intensity, and *reciprocal ties* on the social media platform, where I precisely identify peer groups.

I analyze activism surrounding the abortion rights debate in Argentina in 2018 and 2020. This debate is great for studying how social interactions shape individual activism for three main reasons. First, the abortion rights debate in Argentina was long-lived. Specifically, Congress debated a bill legalizing abortion on demand twice, in 2018 and 2020 - rejected in the former and passed in the latter. Second, pro-choice and pro-life activists coexisted on and offline; their activism persisted until the law's approval. The differential result of the 2018 and 2020 debates suggests voters' important role in the abortion rights bill's legislative process - as 2019 was an electoral year. Third, not only the political right that originates activism - abortion rights - is controversial and normative, but also actions are observable.[4] Altogether suggest peer activism might influence individuals.

I model social interactions as follows. I conceptualize Twitter as a social network and posting tweets as strategic interactions. Then, I develop a model of heterogeneous peer effects in a network. I assume links connecting activists are of two types: between individuals with aligned or opposing viewpoints on abortion rights. I allow for a differential influence of activist peers depending on the type of link. I do not impose additional assumptions regarding the strategic nature of activism, allowing me to empirically test the existence of substitutability or complementarity in online behavior.

The model estimates reveal the existence of strategic complementarity in online activism. Notably, this strategic complementarity comes from both aligned and opposing activist peers. The evidence suggests that the composition of the peer group plays a role in understanding individual activism. Remarkably, the exposure to *early activism*, approximated by the proportion of peers who were activists before the first Congress debate in 2018, is associated with a higher strategic complementarity, but only for like-minded activists. Early activism speaks to the tenure of peers' activism

[4] That is, individuals can observe the activism of their peers.

- if they are persistent activists or newcomers. As such, I interpret this result as a differential impact of activism tenure, depending on the link type, i.e., connecting aligned or dissident peers.

To conduct the empirical analysis, I recover Twitter's network where online activism happens. I build a longitudinal dataset of Twitter users where ties and actions are observable. The construction of this novel dataset involves two significant challenges - determining online activism and identifying social media users engaged in the abortion rights debate to recover their network. For the former, I define *online activism* as the product of two terms: its *intensity* - the daily count of abortion-related tweets posted by any user - and its *sign* - pro-choice or pro-life. My approach for the latter is defining an *initial node* of Twitter's network as any user who has posted at least one abortion-related tweet during each Congress debate - in 2018 *and* 2020. For any initial node, I define her *peer group* as the set of users who follow and are followed by the user - her reciprocal ties. In addition, I download the mutual ties of a randomly selected one percent of her peers - which I name *peers-of-peers.*

I find suggestive evidence of *homophily* in the formation of Twitter's network. Homophily is a tendency to interact with similar individuals - along many dimensions of similarity. In this paper, I find that abortion-rights activists, either pro-choice or pro-life, are highly connected through Twitter - on average, 24% of the users in the peer group are also activists. Nonetheless, the evidence does not support the hypothesis of an *echo-chamber effect*, i.e., the segregation of individuals into like-minded groups, which induces polarization as they interact together. First, for most users, there is no chamber - on average, two-thirds of the activist connections share views on abortion rights, but the remaining one-third are dissidents. Second, there is no echo - the peer effects estimates for like-minded activists do not vary for users with relatively more homogeneous or heterogeneous peer groups.

Following the empirical literature on peer effects, my identification strategy relies on the *partially overlapping network's property.* This property relates to peer groups being individual-specific when social interactions are structured through a network. Bramoullé et al. (2009) and De Giorgi et al. (2010) have shown that this feature helps identify peer effects, as indirect links are a source of valid instrumental variables for peers' actions. Then, I propose a network-based instrumental variable to estimate the parameters. As Twitter data does not provide detailed individual characteristics, I take advantage of the longitudinal structure of the data to include individual fixed effects, which allow me to control unobserved factors driving individuals' actions and network formation.

Related Literature. This paper contributes to the empirical understanding of the social motives of political activism and collective action problems. Cantoni et al. (2019) and Hager et al. (2023) highlight the role of beliefs about others' protest turnout on individual participation, finding strategic substitutability in protest behavior. Enikolopov et al. (2020) show that social image plays a role in the decision to participate in a protest. They also find that online and offline protest participation is positively associated. Closer to my paper, González (2020) finds strategic complementarity in the protest behavior of Chilean students - pointing out a coordination mechanism, and Bursztyn et al. (2021) identify that social interactions are crucial for sustained political engagement. However, their observation of individual networks is approximated by high school and university classmates, respectively. This paper complements the previous studies (i) by providing a precise observation of peers as Twitter links and (ii) by focusing on the intensive rather than the extensive margin of political activism.

This paper also speaks to the empirical literature on peer effects[5] - which has found evidence of strong effects in different aspects of life: education (Patacchini et al. (2017), De Giorgi et al. (2010)), female labor supply (Nicoletti et al. (2018)), financial decisions (Bursztyn et al. (2014)), and consumer behavior (Moretti (2011), De Giorgi et al. (2020)), among others. Nonetheless, the study of heterogeneous peer effects is usually overlooked - where this heterogeneity refers to a differential response of individuals to different types of peers. A relevant exception is Patacchini et al. (2017), which estimates heterogeneous peer effects in education. Relying on the National Longitudinal Survey of Adolescent Health data, the authors differentiate the peer influence by the tenure of the links and find a persistent peer effect for long-lived links. Consistently with my case study, the source of heterogeneity in the link types relates to the users' viewpoint on abortion rights. Additionally, in this paper, I provide novel evidence on the role of peer effects on social media platforms; in a context where activism is closely related to political rights and social norms.

Lastly, this paper contributes to understanding who - and how individuals - engage in online social interactions, especially in the political sphere.[6] Halberstam and Knight (2016) study the type of links that politically engaged users form, finding homophily in their Twitter network. Nonetheless, Gentzkow and Shapiro (2011) reveal that online interactions are less segregated than offline. Conover et al. (2011) show that political retweets are highly segregated along partisan lines, but user mentions are not - as dissidents mention each other frequently. Larson et al. (2019) find that Charlie Hebdo protest participants were more connected to each other through Twitter when compared to users who did not participate. I consider social ties as reciprocal links on Twitter and study a political right without a partisan position in the Argentinian context. Regarding the proportion of like-minded and dissident peers, the data reveals heterogeneity in the peer group composition - pointing out that some users are segregated into like-minded groups, but the majority are not.

The rest of the paper is organized as follows. Sub-section 1.1.1 introduces the study case. Section 1.2 presents the theoretical framework, and section 1.3 describes

[5] See Bramoullé et al. (2020) for a review.

[6] For a review, see Zhuravskaya et al. (2020).

the data. Sections 1.4 and 1.5 present the peer effects estimates and the robustness checks, respectively. Section 1.6 concludes.

1.1.1 Abortion rights and activism in Argentina

In December 2020, the Argentine Congress legalized abortion on demand. Nevertheless, it was not the first time the Argentine Congress studied that bill. Before that successful attempt, pro-choice activists had put forward the same bill in Congress seven times - from 2005 onward. The legislative branch in Argentina is bicameral, consisting of a Senate and a Chamber of Deputies. A bill put forward by a popular initiative has to go through three steps to become law. First, a subcommittee of the Chamber of Deputies receives it. The subcommittee has up to two years to send the bill to the Chamber of Deputies. If that happens, deputies study the bill. Finally, the Senate debates it. If both cameras pass the bill, it becomes law.

In 2018, the abortion rights bill reached Congress for the first time. Before, it never went further than the deputies' subcommittee. The Chamber of Deputies passed the bill in June 2018. However, In August 2018, the Senate rejected it by a low margin. As mentioned before, both cameras finally approved the law in December 2020. The previous year, 2019, was an electoral year in Argentina. As a result of the national elections, one-third of the seats in Congress changed. Even though abortion rights was a non-partisan topic, most candidates' statements included their position. Moreover, Congress members on seats in the two debates, 2018 and 2020, did not change their votes. This evidence suggests voters' important role in the abortion rights bill's legislative process.

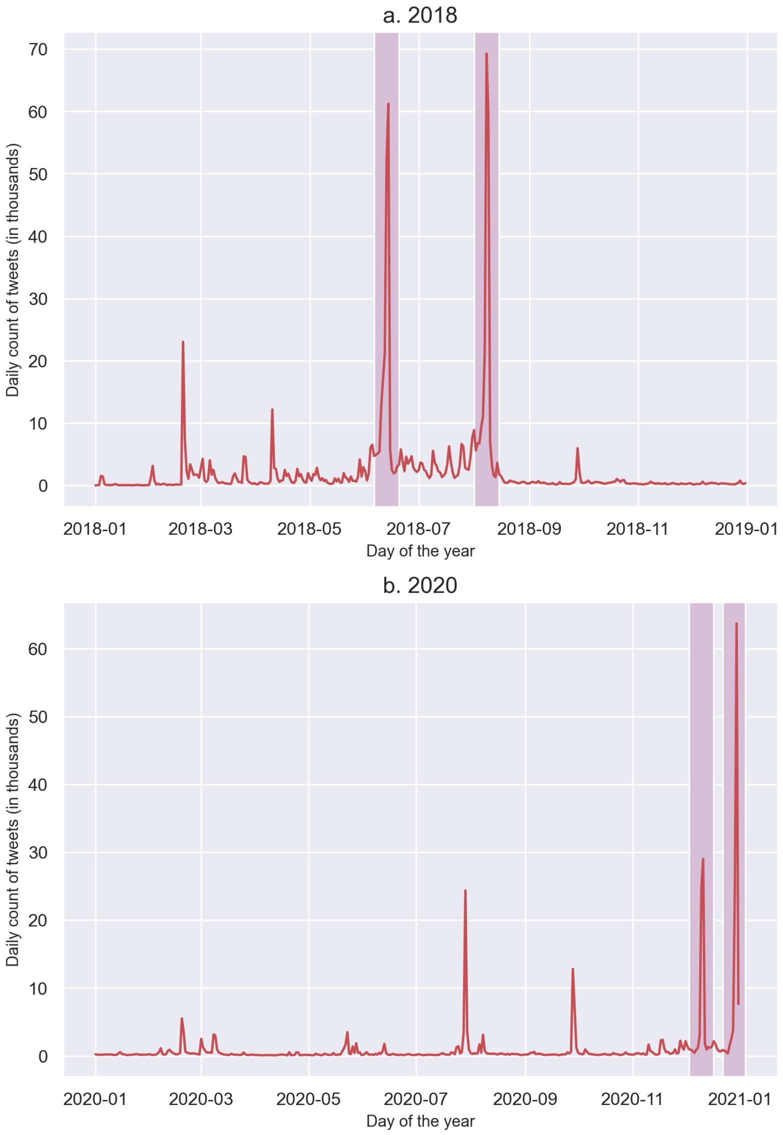
Abortion rights were, and still are, a controversial aspect of reproductive rights in Argentina. Yet, this is not particular to Argentina but common to many Latin American countries - where abortion access is restrictive.[7] The first evidence of this is the difficulty passing the law. The sustained mobilization of pro-choice activists and their counter-mobilization by pro-life activists constitutes the second piece of evidence. Pro-choice and pro-life activists organized many public demonstrations over the period. Furthermore, they designed two handkerchiefs to signal their advocacy, which crossed the Argentine borders and became a symbol of abortion rights mobilizations.[8] Crucially, the *online presence* of pro-choice and pro-life activists - the focus of this research - was vigorous. Figure 1.1 shows the daily count of abortion-related tweets in 2018 and 2020. Twitter activity peaks coincide with days

[7] Although they have different abortion regulations, most restrict abortion access, and only a few allow on-demand abortions. More information is at this link.

[8] Figure 1.5 in Appendix shows these two bandanas, green for the pro-choice activists and light-blue for pro-life activists. Media pictures of these handkerchiefs are found in abortion rights mobilizations in other Latin American countries and the public demonstrations of Roe vs. Wade in the U.S.

when Argentine Congress debated the bill.[9]

Figure 1.1 Abortion-related tweets in 2018 and 2020



Note: Daily count of abortion-related tweets, net of retweets, in the two years of debate, 2018 and 2020. Shadow areas indicate weeks of legislative debate on the abortion bill.

1.2 Theoretical framework

I study a model of social interactions, where individuals choose their level of involvement in *online activism* related to a specific topic A in a *predetermined network.* The links between individuals included in the network represent mutually beneficial relationships. Importantly, individuals care about the activism of individuals with whom they interact. Activism is characterized by its *intensity* and *sign.* Intensity

[9] Specifically, days with legislative activity were June 13th and August 8th, 2018, and December 10th and 29th, 2020.

relates to the individual effort in devoting time to being an activist. The sign of activism denotes whether the individual is an activist for or against cause A.

1.2.1 A model of peer effects in a network

Consider an online platform comprised of $n < \infty$ individuals, where $N = \{1, ..., n\}$ is the set of individuals. Each user i has a specific peer group, P_i of size n_i. Let g be the network representing online links between those individuals, and $G = [g_{ij}]$ be the $n \times n$ non-negative adjacency matrix. The (i, j) entry of G, denoted g_{ij}, equals $1/n_i$ if individuals i and j have a link and zero otherwise. I normalize diagonal elements of G to zero so that $g_{ii} = 0 \quad \forall i \in N$. To capture meaningful online links, I assume the network g is undirected, i.e., $g_{ij} \neq 0$ if and only if $g_{ji} \neq 0$.[10]

Conditional on the network structure and their preferences, individuals choose online activism, denoted by $a_i \in (-\infty, \infty)$. Importantly, $|a_i|$ denotes the intensity of activism, and the sign of a_i indicates whether i is for or against A. Each individual i has an ideal point of online activism, denoted $\theta_i \in (-\infty, \infty)$. Since the nature of interactions between individuals with equal-sign and opposite-sign ideal points may differ, I decompose the adjacency matrix G into two matrices, $H = [h_{ij}]$ and $K = [k_{ij}]$. Specifically, the matrix H includes all links in G between individuals of equal-sign ideal points, whereas K does it for opposite-sign. Thus, for any entry (i, j) of the matrices H, K, and G, the following hold:

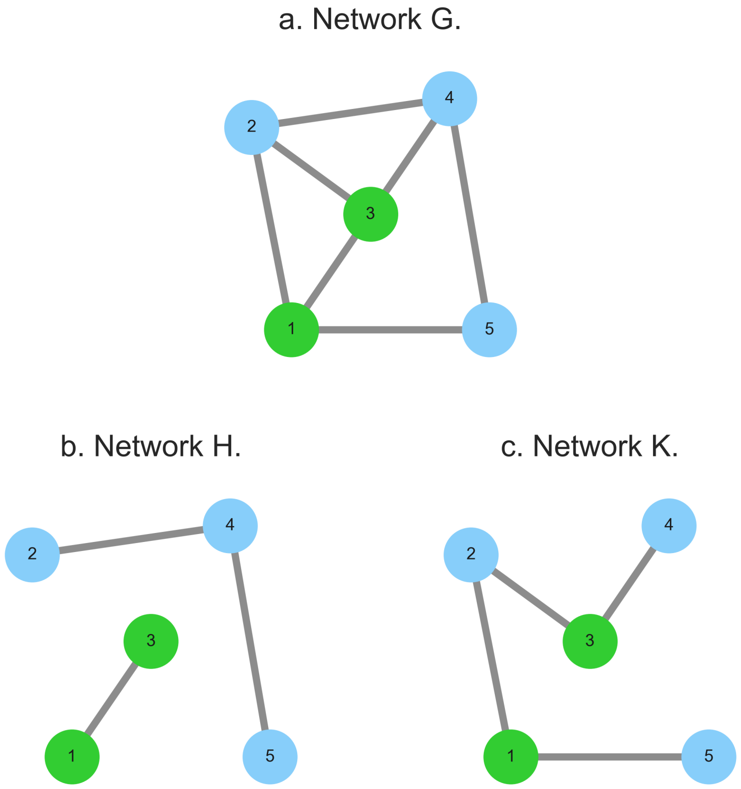
$$\begin{aligned} h_{ij} &\equiv \mathbb{1}_{\theta_i \times \theta_j > 0} g_{ij} \\ k_{ij} &\equiv \mathbb{1}_{\theta_i \times \theta_j < 0} g_{ij} \\ G &\equiv H + K \end{aligned}$$

Figure 1.2 exemplifies the adjacency matrix G decomposition into the matrices H and K. Panels a, b, and c show the network representation of the matrices G, H, and K, respectively. In this example, $N = \{1, 2, 3, 4, 5\}$, $\theta_i > 0$ for $i \in \{1, 3\}$ (green nodes), and $\theta_i < 0$ for $i \in \{2, 4, 5\}$ (blue nodes). Thus, the network representation of H only includes links *within* the subsets $\{1, 3\}$ and $\{2, 4, 5\}$ whereas the network representation of K includes links *between* those subsets.

Following the literature, e.g., Ballester et al. (2006), Bramoullé et al. (2014), I assume a linear quadratic specification for the utility of activism levels. Considering that activists for or against topic A may interact differently, the model allows for heterogeneous peer effects. The parameter β reflects peer effects when the sign of own and peers' ideal points coincide, i.e., individuals whose link belongs to matrix H. In contrast, γ measures peer effects when it differs, i.e., individuals whose link belongs to matrix K. Throughout this paper, I assume that $|\beta| < 1$ and $|\gamma| < 1$.

[10] In the empirical section, I check the sensitivity of the results to this assumption. Section 1.5 discusses it.

Figure 1.2 Decomposition of the adjacency matrix G



Note: Links between individuals with aligned viewpoints on topic A belong to matrix H, whereas links between individuals with opposing viewpoints on topic A belong to matrix K.

Denoting any profile of activism levels by $\mathbf{a}$, the following function represents i's utility:

$$u_i(\mathbf{a}, G) = u_i(\mathbf{a}, H, K) = \theta_i a_i - \frac{1}{2}a_i^2 + \beta \sum_{j \in N} h_{ij} a_i a_j - \gamma \sum_{j \in N} k_{ij} a_i a_j \tag{1.2.1}$$

The first two terms of equation (1.2.1) reflect i's private benefit and cost associated with her activism level. The third and fourth terms represent the heterogeneous social benefit or cost of changing an individual's action. As activism signs differ for two peers with opposing viewpoints, I include the second term of social interactions preceded by a negative sign. This modeling choice allows me to interpret the strategic nature of activism in the usual manner, i.e., a positive parameter reflects complementarity, whereas a negative substitutability. Individuals play a non-cooperative game for the choice of the activism levels, conditional on the network structure. The equilibrium concept is Nash equilibrium. For any individual i, the best-response function is given by:

$$a_i^{BR} = \theta_i + \beta \sum_{j \in N} h_{ij} a_j^{BR} - \gamma \sum_{j \in N} k_{ij} a_j^{BR} \tag{1.2.2}$$

Denoting the ideal points vector by θ, the system of best-response functions in matrix notation equals:

$$\mathbf{a} = \theta + \beta H \mathbf{a} - \gamma K \mathbf{a} \tag{1.2.3}$$

Provided $|\beta| < 1$ and $|\gamma| < 1$, $[I - \beta H + \gamma K]^{-1}$ exists, where I is the $n \times n$ identity matrix, the equilibrium is determined as:

$$\mathbf{a}(H, K) = [I - \beta H + \gamma K]^{-1} \theta \tag{1.2.4}$$

In Appendix 1.A.1, I prove the condition for the invertibility of $[I - \beta H + \gamma K]$ and comment on the equilibrium uniqueness.

1.2.2 Discussion and extensions

Despite its simplicity, the model captures the following essential aspects of online interactions: (i) the network structure of social media platforms like Twitter, (ii) the interdependency between individuals' actions, and (iii) the potential heterogeneity in peer effects. In addition, the model is suitable for the empirical estimation of these heterogeneous peer effect parameters, which constitutes one of the main objectives of this project. Patacchini et al. (2017) also estimates heterogeneous peer effects in education, differentiating the parameters by the tenure of the links, i.e., long-lived vs. short-lived links. Consistently with my case study, the source of heterogeneity of peer effects in the model relates to the individuals' viewpoint on topic A.

According to the model predictions, any individual's activism level is a weighted sum of her preferences, θ_i, and the average activism levels of her peers. If the social connections were irrelevant to explaining activism, the optimal solution for any i is simply $a_i^* = \theta_i$. Social interactions matter if at least one parameter (β, γ) differs from zero. A positive value on the peer activism parameters, β and γ, indicates strategic complementarity in the intensity of activism, while a negative value indicates substitutability. Even though activism of dissident peers has, by construction, opposed signs, the interpretation of γ is the traditional one - as a negative sign precedes the parameter in the utility function.

A limitation of this model is the assumption that the network is predetermined. In that sense, a possible extension would explicitly study network formation[11] in addition to the strategic interactions. In that case, the game would be a two-stage game. Individuals first form their online social network and then choose their level

[11] See De Paula (2020) for a review of econometric models of network formation.

of involvement in online activism. Taking equation (1.2.1) as a reference, the utility for i would be given by:

$$u_i(\mathbf{a}, G) = \theta_i a_i - \frac{1}{2}a_i^2 + \beta \sum_{j \in N} h_{ij} a_i a_j - \gamma \sum_{j \in N} k_{ij} a_i a_j + \sum_{j \in N} g_{ij} \psi(i, j)$$

The fifth term denotes i's explicit preferences over the online network structure. The function $\theta(i, j)$ determines how much i values j as a peer in the network. It can depend on different variables, including i's preferences for her and j's degree, a measure of common interests, among others - see, for example, Hsieh et al. (2020). A different approach for network formation would be the one proposed by Goldsmith-Pinkham and Imbens (2013) and Hsieh and Lee (2016). The network formation process is modeled via pairwise stability,[12] while the outcome is specified following equation (1.2.4).

1.3 Data

My primary data source is the platform Twitter. I aim to understand how social interactions affect online activism, considering that these interactions could happen between users with aligned or opposing viewpoints in the abortion rights debate. I need to construct a dataset with observable actions and links. In that respect, the first challenge is determining *what online activism is.* In the empirical analysis, I consider online activism as the number of abortion-related tweets posted by a user in a given period. Then, to measure it, the first step is building a *tweets' dataset.*

The second challenge is identifying social media users engaged in the Argentinian abortion rights debate. Given that Twitter is a giant online network, I need to restrict my attention to a sub-sample of users to conduct the empirical analysis. My approach is to define the *initial nodes* of the network as the set of users who fulfill specific requirements. Then, by identifying these users, I construct the Twitter network where online activism is happening, which I name the *users' dataset.* I create a panel dataset with an explicit network structure by combining the tweets' and users' datasets. The following paragraphs explain how I build and merge these two datasets.

To create the tweets' dataset, I first collect the set of abortion-related tweets from 2010 to 2020. I download all the tweets that contain at least one abortion-related hashtag.[13] Twitter activists popularly used these pro-choice and pro-life hashtags to express their opinion. Further, activism through Twitter is often associated with specific hashtags, as documented in the literature Jackson et al. (2020). The tweets' dataset includes all the replies and quotes to any of those tweets but excludes

[12] Jackson and Wolinsky (1996), Calvó-Armengol and Ilkılıç (2009), Jackson and Watts (2001).

[13] Table 1.6 in Appendix provides the list of hashtags used in the Twitter query.

retweets. I exclude them because of the noise they introduce in classifying pro-choice and pro-life tweets. First, "*retweet* $\neq$ *endorsement*" is widespread on Twitter. Second, retweeting is a Twitter action of low stakes compared to posting or replying to tweets - but their quantitative comparison is not trivial.[14]

I filter tweets according to their content and the account that posted them. The filtering criteria select Twitter accounts (i) with a positive number of links and (ii) which are not news outlets, organizations, or trending-topic trackers, among others. I further restrict the dataset to (i) tweets in Spanish and (ii) which do not correspond to an abortion rights debate in another country where Spanish is an official language. Moreover, in the empirical analysis, I restrict my attention to the years 2018 and 2020 for two reasons. This period concentrates most of the tweets. Additionally, it coincides with when the Argentine Congress debated the abortion rights bill. The final tweets' dataset includes 2 million observations.

The primary variable of interest, named *online activism* and denoted by a_i, is the product of two terms. Activism intensity, as the daily count of abortion-related tweets posted by any user, $|a_i|$; and activism sign, stating whether she is a pro-choice or pro-life activist.[15] Following this procedure, I compute an integer-valued variable $a_i \in \{..., -2, -1\} \cup \{1, 2, ...\}$. I assign the value $a_i = 0$ for any user on the dates she did not post an abortion-related tweet. In that way, activism is an integer-valued variable in the interval $a_i \in \{..., -1, 0, 1, ...\}$.

To determine the activism sign, I need to classify all the tweets posted by a user on a given day as pro-choice or pro-life. To accomplish this, I proceed as follows. First, I classify a tweet as pro-choice (pro-life) if it only contains pro-choice (pro-life) hashtags. Then, I use a series of tuples of words to refine this classification. For example, suppose a tweet includes the hashtag *"#AbortoLegal"* - legal abortion, in Spanish - and *"feminazi"* - the combination of feminist and Nazi. In that case, I classify it as a pro-life tweet. Finally, I compute the average activism sign per day and individual and reclassify tweets to match the sign of this mean. This last step implicitly assumes individuals do not change their opinions in a short period, in this case, a day. Importantly, this procedure categorizes users into pro-choice and pro-life activist groups daily, allowing users to switch positions over more extensive periods. Nonetheless, I do not observe users switching between one and another movement.

Second, I construct the online network of Twitter users engaged in the Argentinian abortion rights debate, which I previously named the users' dataset. The first step is to define the *initial nodes* of the network. I consider as an initial node any user who fulfills the following conditions: (i) the user has posted at least one

[14] Conover et al. (2011) suggest retweeting is a Twitter action that goes along partisan lines. Thus, if any, by not considering retweets, I am computing a lower bound of online activism.

[15] A 90% of the initial nodes are pro-choice activists, whereas the 10% remaining is composed of pro-life activists.

abortion-related tweet during the Congress debates in 2018 *and* 2020, (ii) she has less than 5.000 connections on Twitter, and (iii) the user provides geo-location information.

The upper bound imposed on connections works twofold. First, it limits the possibility of including celebrities, influencers, and politicians in the users' dataset. The theoretical model presented in Section 1.2 may be unsuitable for these individuals as their incentives could differ from the rest of Twitter users. For instance, politicians' tweets could obey their perceived probability of being elected, and celebrities may decide not to express their opinion to preserve their public image. Additionally, I impose this restriction for tractability.[16]

After applying this filtering criterion, the users' dataset contains approximately 6.000 initial nodes. For any initial node, I download a list of her mutual connections, i.e., an account that follows and is followed by that user. I define these users as *peers* in the empirical analysis. I restrict my attention to reciprocal links to recognize the different natures of unilateral and bilateral relationships. Finally, I download the list of mutual connections for randomly selected one percent[17] of the peers in the network. I name them *peers-of-peers*. These three types of users, initial nodes, peers, and peers-of-peers, form the users' dataset.

For consistency, I filter accounts with less than 5.000 connections for peers and peers-of-peers. Furthermore, I only keep Twitter accounts whose creation date is 2018 or earlier. This condition is crucial, given how the Twitter API works. Its *follows-lookup endpoints* return connections on the day the request is made.[18] Therefore, it is impossible to observe the Twitter network for a given time in the past. Applying the filtering criterion of creation date, I approximate the observed network as much as possible to the 2018-2020 network.

Additionally, I classify users according to their participation in abortion rights activism into three groups. A user is *non-activist* if she does not appear in the tweets' dataset. She is an *activist* if she appears in the tweets' dataset at least once and an *early activist* if she appears before the first Congress debate in June 2018. For any user i in a given day t, her *activity status* could be *t-posting* or *t-not-posting*, depending whether $a_{it} > 0$ or $a_{it} = 0$. Thus, a non-activist is a user whose activity status equals t-not-posting for all the periods. At the same time, an activist is any user whose activity status equals t-posting at least for some t. Thus, the categorization of peers into activist or non-activist is time-invariant, whereas the activity status of activist peers depends on the specific date t.

Table 1.1 summarizes initial nodes' degree distribution, i.e., their peer group size, and its decomposition into the categories of activists and early activists. On

[16] Twitter is a giant network, so restricting the number of connections alleviates the computational burden.

[17] For each i, I download that list for the closer natural number to $1\% n_i$.

[18] Twitter requests to this endpoint were made between December 2021 and February 2022.

average, individuals have 412 reciprocal links, of which 97 are activists. Moreover, the set of activists who were t-posting on a given date t, of size n_{it}^{A}, is a subset of the set of activists among peers, of size n_{i}^{A}. The latter category is the *relevant* in the model estimation. The last column of Table 1.1 reports that, on average, across time and individuals, initial nodes have 20 t-posting peers. Combined with the full observability of Twitter links, small peer groups make this context ideal for studying peer effects.

Table 1.1 Initial nodes' degree

	n_i	n_i^A	n_i^{EA}	mean n_{it}^A
Mean	412	97	45	20
St.Dev.	509	142	65	30
Min.	2	1	0	0
Median	250	45	19	7
Max.	4612	1723	805	378
Individuals				5808

Note: n_i denotes the size of the peer group, whereas n_i^A and n_i^{EA} is the size of the peer group classified as activists and early activists, respectively. mean n_{it}^A is the mean size, across time, of activist peers who were t-posting at t.

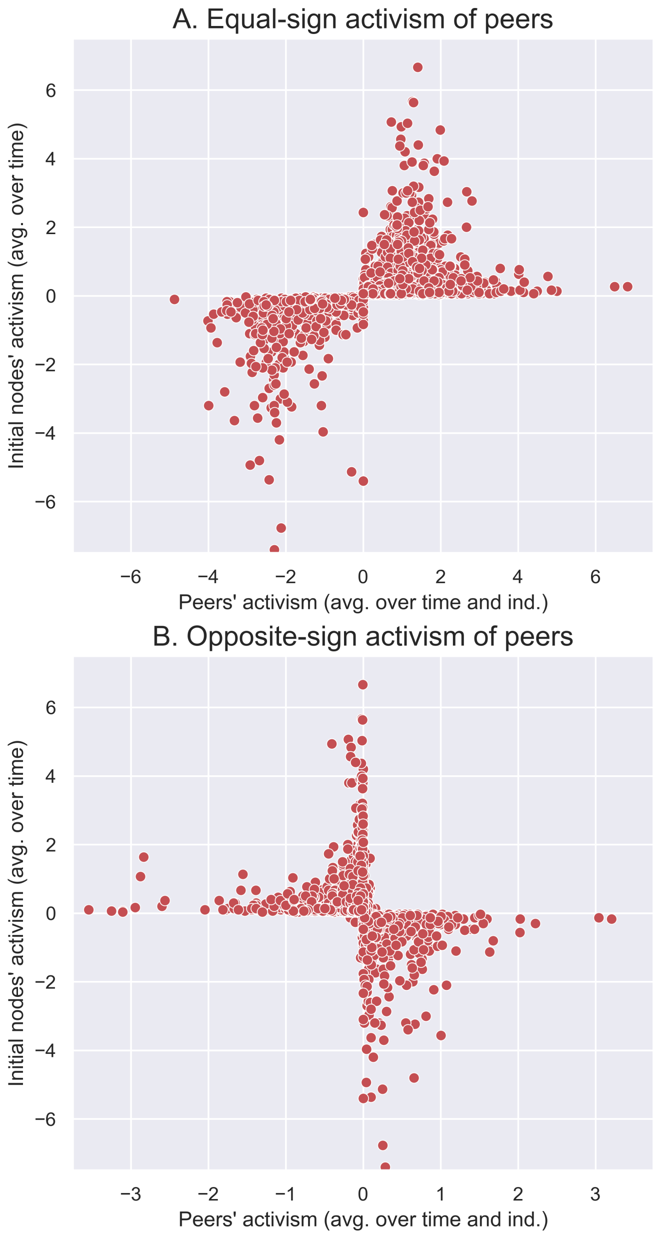
Finally, I combine the tweets' and users' datasets previously mentioned to generate a *panel dataset* with an explicit *network structure*. For any initial node, I observe (i) the set of her first-degree connections, (ii) a sub-set of her second-degree connections, and (iii) the value of online activism for her and her observable connections. The panel dataset is balanced, each individual is an initial node, and the period is a day. In the empirical analysis, I use the dataset with observations for a one-week window centered on each day Congress debated the abortion rights bill.

1.3.1 Descriptive statistics

Figure 1.3 presents correlations between initial nodes' activism and the average activism of their peers. The variable on the x-axis is the average of peers' activism over time and per individual. On the y-axis, the variable is the average over time of the initial nodes' activism. Panel A illustrates it for equal-sign peers' activism, whereas Panel B is for opposite-sign peers' activism.

While the correlation between equal-sign peers' and own activism is positive, its analogous statistic for opposite-sign activism is negative. Since the intensity of activism is its absolute value, the sign of the two correlations reflects a positive relationship between activism intensities. The intensity of pro-choice (pro-life) activism increases as it becomes more positive (negative). Therefore, a more intense opposite-sign peers' activism correlates positively with higher own activism.

Figure 1.3 Correlation between initial nodes' and peers' average activism.



In the two panels, points in which activism of initial nodes is close but not equal to zero reflect that the user was t-not-posting on Twitter for some of the dates considered in the empirical analysis. Thus, the source of variation in initial nodes' activism is twofold: the intensity of their activism on the days they were t-posting and the frequency of that activity status.

There is a notable difference between Panel A and B of Figure 1.3. While in Panel A, there are a few points in which peers' activism is close to zero, in Panel B, those points correspond approximately to a third of the total number of initial nodes.[19] In other words, a third of the users considered as initial nodes do not have links with users whose (average over time) activism has the opposite sign. Moreover,

[19] 1636 out of the 5808 users.

this is true for initial nodes participating in both movements, pro-choice and pro-life. Nonetheless, two-thirds of the individuals are connected to users with opposing and aligned viewpoints on abortion rights. I interpret this as evidence against the existence of an *echo chamber*. A necessary condition for this phenomenon is the existence of a chamber: the segregation of users into like-minded groups.

Table 1.2 Descriptive statistics

	Mean	Median	Std. Dev.
Panel A: Pro-choice initial nodes			Ind. 5225
activism	0.305	0.167	0.450
peer activism$_{\text{equal-sign}}$	0.803	0.736	0.515
peer activism$_{\text{opposite-sign}}$	-0.067	-0.021	0.175
t-posting peers$_{\text{equal-sign}}$	13.460	4.733	20.447
t-posting peers$_{\text{opposite-sign}}$	6.193	2.000	9.605
Panel B: Pro-life initial nodes			Ind. 583
activism	-0.653	-0.300	1.055
peer activism$_{\text{equal-sign}}$	-1.388	-1.283	1.113
peer activism$_{\text{opposite-sign}}$	0.321	0.200	0.386
t-posting peers$_{\text{equal-sign}}$	15.037	5.300	25.844
t-posting peers$_{\text{opposite-sign}}$	6.638	2.733	9.968
Panel C: all initial nodes			Ind. 5808
activist peers$_{\text{ratio}}$	0.237	0.207	0.159
early activist peers$_{\text{ratio}}$	0.448	0.455	0.165
activist peers-of-peers$_{\text{ratio}}$	0.158	0.132	0.123
early activist peers-of-peers$_{\text{ratio}}$	0.419	0.426	0.169

Note: Panel A and B variables in this table are averaged over time and individuals, whereas Panel C variables are averaged over individuals. The activist peers ratio is the proportion of activists in the peer group. The early activist peers ratio is the proportion of early activists among activist peers.

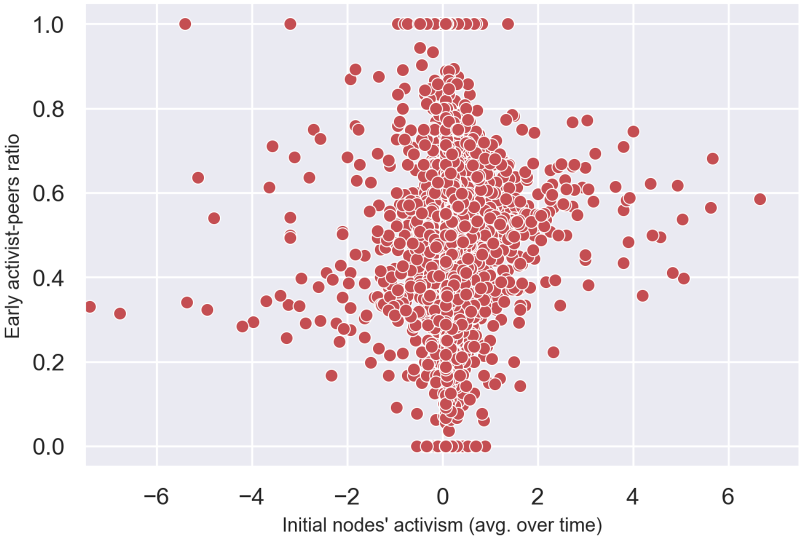
In this line, Table 1.2 presents complementary information. In Panels A and B, I summarize the main variables of the model, averaged over time. They include initial nodes' and peers' activism and the number of t-posting peers. Panel A corresponds to the initial nodes classified as pro-choice activists, while Panel B does it for pro-life activists. Lastly, Panel C presents descriptive statistics of the ratio of activist and early activist users in the peer groups and among peers-of-peers. The mean of all the activism variables differs from zero over time and by individuals. Consistently with Figure 1.3, opposite-sign activism is the variable whose mean is closer to zero. On average, pro-choice initial nodes have 13 pro-choice and 6 pro-life t-posting peers per day. For pro-life initial nodes, these numbers are 15 and 7. Therefore, around two-thirds of peers are like-minded activists, whereas one-third are not.[20]

[20] Note that t-posting peers$_{\text{equal-sign}}$ and t-posting peers$_{\text{opposite-sign}}$ is the decomposition of the

According to Panel C, 24% of users in the peer groups are activists, on average.[21] The information in Table 1.2, jointly with Figure 1.3, suggests that users engaged in the abortion rights debate are highly connected but not perfectly polarized into two groups. While the literature studying the existence of online echo chambers is inconclusive,[22] there is evidence that activists are highly connected through social media, e.g., Larson et al. (2019). Accordingly, the description of this context is consistent with *homophily* in Twitter's network, in the sense of being engaged in the abortion rights debate but not necessarily sharing viewpoints.

Finally, Figure 1.4 presents the correlation between initial nodes' activism and the ratio of early activists in her peer group. On the x-axis, the variable is the average over time of the initial nodes' activism. The y-axis variable is the proportion of early activist peers over the number of activist peers. According to Table 1.2, on average, 24% of peers are activists, and 45% among those are early activists. As Figure 1.4 shows, the differential exposure of initial nodes to early activism is a source of variation in the data (at the individual level). Significantly, the exposure to early activism varies for both pro-choice and pro-life initial nodes. As mentioned above, I define an early activist as any user who appears in the tweets' dataset before June 2018, the month of the first Congress debate on the abortion rights bill. In that regard, I interpret early activism as a measure of persistence, even strength, in online activism. Therefore, differential exposure to early activists may play a role in explaining peer effects.

Figure 1.4 Correlation between activism and the early activist-peers ratio.



total amount of t-posting peers, which is reported in the last column of Table 1.1 as n_{it}^{A}, but without differentiating between pro-choice and pro-life initial nodes.

[21] Appendix 1.A.3 provides further information and descriptive statistics, including the histograms of the variables in Table 1.2.

[22] See Levy and Razin (2019) for a review of echo chambers.

1.4 Empirical analysis

In this section, I follow an instrumental variables approach to estimate the heterogeneous peer effect parameters. Consistently with section 1.2, I estimate peer effects by contemplating links between like-minded users and users with opposing viewpoints on abortion rights. The identification strategy relies on the *partially overlapping network's property*, which allows me to propose network-based instruments. In addition, and taking advantage of the longitudinal data structure, I include individual fixed effects to control for unobserved factors driving online activism and network formation.

Before discussing the identification strategy, a clarification is relevant. I estimate peer effects for a sub-sample of the Twitter population: those who posted abortion-related tweets during the legislative debates on the bill. Extrapolating the results of the estimation in this study to the entire Twitter population would require assuming that the peer parameters among users who participate and who do not participate are equal. In other words, I estimate peer effects on the *intensive margin* of online activism. Although interesting, the estimation of peer effects on the participation decision, i.e., the *extensive margin* of activism, is out of the scope of this paper. That estimation would require detailed individual characteristics[23] as well as the observation of the entire Twitter population.

1.4.1 Estimation and identification

It is a well-known challenge in the peer effects literature to disentangle the mechanisms behind the interdependence-in-actions of individuals who interact together. In his seminal paper, Manski (1993) distinguishes three sources of this interdependence: contextual, endogenous, and correlated effects. The *contextual or exogenous effect* is the influence of exogenous peers' characteristics on an individual's actions. The *endogenous peer effect* is the impact of peers' actions on an individual's actions. Lastly, individuals and their peers may behave similarly due to sharing a common environment, the so-called *correlated effect.* Therefore, the causal estimation of endogenous peer effects requires disentangling them from contextual and correlated effects. This distinction becomes easier when interactions are structured through a network.

When a network structures social interactions, the peer group of any individual is *specific* to her. This feature alleviates Manski's *reflection problem*, making the distinction between endogenous and exogenous effects possible. Specifically, the reflection problem is a consequence of the simultaneity in the behavior of individuals, see equation (1.2.3), and it arises only under the assumption of group-wise

[23] Matching Twitter data with other data sources at the individual level is against Twitter Developer Account's terms and conditions.

interactions.[24] Even though I do not estimate exogenous effects and, instead, I control for them by using individual fixed effects, network data is still crucial for the identification strategy. The reason is the (potential) existence of correlated effects, that is, group-specific unobserved variables driving individual's and peers' actions. Since peer groups are individual-specific, the characteristics of indirect links in the network are valid instrumental variables for peers' actions.[25]

In this paper, I follow a network-based instrumental variable approach to causally estimate peer effects.[26] Specifically, I rely on the *partially overlapping network's property* to estimate the peer effects parameters, see Bramoullé et al. (2009) and De Giorgi et al. (2010). Given that individuals interact in a social network, two connected individuals, i and j, have different peer groups, P_i and P_j. Importantly, the existence of *intransitive triads* helps to identify peer effects. An intransitive triad between individuals (i, j, l) exists if, for the pair of individuals (i, j), there exists an individual l connected to j but not to i. In simple words, from i's perspective, l is a friend of her friend, j. Formally,

$$i \in P_j \quad \text{and} \quad l \in P_j \quad \text{but} \quad l \notin P_i$$

For any individuals i and j, I define $P_{j/i}$ as the set of individuals l who form intransitive triads with them. If i is an initial node and j is her peer, I use individuals on the set $P_{j/i}$ to instrument for peers' activism. As I estimate heterogeneous peer effects, I split this set and the peer group P_i into two subsets each: $(P_i^H, P_{j/i}^H)$, containing information about equal-sign activism, and $(P_i^K, P_{j/i}^K)$, about opposite-sign activism. The proposed instrumental variables are the *daily ratios of equal-sign and opposite-sign t-posting users* among those in $(P_{j/i}^H; P_{j/i}^K)$. Given the available data, the following remark is essential. The instrument is the activity status of the peers of a 1% randomly selected sample of initial nodes' peers. That is, I observe the activity status from users included in the sets $P_{j/i}$ from a 1% of the peers $j \in P_i$. For a given date t and initial node i, I compute the ratio of equal-sign and opposite-sign t-posting users as the proportions of those in the union of the observed sets, $P_{j/i,t}$.

To gain intuition about the identification strategy, recall the ratios of equal-sign and opposite-sign t-posting users on the sets $P_{j/i,t}$ measure the daily exposure of peers $j \in P_i$ to online activism. The construction of these ratios depends on the randomly assigned observability of the sets $P_{j/i}$, generating an additional source of variation. Then, the observed ratios measure the exposure to online activism of 1% randomly selected peers $j \in P_i$. The identifying assumption is, therefore, that the

[24] That is, when individuals are affected by all individuals belonging to their group and by nobody outside them.

[25] The indirect links of any individual share a common environment with the individual's peers but not with her.

[26] In the context of Twitter, Cagé et al. (2022) also use a network-based instrument to study the information propagation from social media to mainstream media.

activity status of the observed peers-of-peers, $l \in P_j$, who are not directly connected to an initial node, $l \notin P_i$, only affects her activism, a_i, through the activism of peers, $j \in P_i$.

For any initial node i and day t, the parametric specification of the individual heterogeneity θ_{it} and the resulting empirical counterpart of equation (1.2.2) are:

$$\theta_{it} = \theta_x x_{it} + \theta_{LD} + \theta_i + \epsilon_{it}$$
$$a_{it} = \theta_x x_{it} + \beta \sum_{j \in P_i^H} a_{jt} + \gamma \sum_{j \in P_i^K} a_{jt} + \theta_{LD} + \theta_i + \epsilon_{it}$$

where x_{it} is a set of covariates related to the tweet's popularity, i.e., the daily average of likes, retweets, quotes, and replies to the user's tweets. θ_i is an individual fixed effect. θ_{LD} is a dummy variable that takes value one when Congress debated the abortion rights bill, i.e., on a legislative day, and zero otherwise. ϵ_{it} is and *i.i.d.* error term with variance σ^2.

Given Twitter data characteristics, including individual fixed effects is crucial for the empirical analysis. Working with social media data has the advantage of clear observability of links but at the cost of lacking detailed individual characteristics, which constitute the source of identifying exogenous peer effects and determining the sorting of individuals into a network. Thus, I include individual fixed effects to control for unobserved factors driving Twitter users' behavior and network formation. The underlying assumption is that such unobserved variables are time-invariant. The empirical literature on peer effects has addressed these threats to identification using network fixed effects. Compared to individual fixed effects, these are less restrictive, for instance, regarding the covariates that can be included in the estimation. In the main specification of the model, I do not include network fixed effects, but in appendix 1.A.3, I show my results are robust to their inclusion.

In the context of social media, a potential threat to identification is given by how the *Twitter algorithm* works. In particular, regarding the content shown in the Twitter feed of any user whose author is not her peer. Although there is no official information about the algorithm, it is reasonable to assume the observation of such content is more likely to happen if the tweet becomes viral or if the tweet's author and the user share connections. Regarding the former, I include tweet popularity measures in the estimation. Finally, the essence of an instrumental variable is that the instrument and the independent variable are related only via the endogenous variable. In the Twitter context, it translates to the user and the tweet's author being related through their peers in common.

1.4.2 Results

Table 1.3 presents the peer effects estimates. Columns (1)-(2) correspond to the Fixed Effects model (FE), whereas Columns (3)-(4) present the results of the instrumental variable approach (IV-FE). Panel A includes all the observations for one-week windows centered on the legislative days,[27] so the panel is balanced. In Panel B, I restrict my attention to observations with non-zero values of initial nodes' activism. In all the specifications, results indicate the existence of complementarities in online activism.

Table 1.3 Peer effects in online activism.

	FE		IV-FE	
	(1)	(2)	(3)	(4)
Panel A: Balanced Panel				
$\text{activism}_{\text{equal-sign}}$	0.194***	0.134***	0.564***	0.376***
	(0.014)	(0.011)	(0.020)	(0.025)
$\text{activism}_{\text{opposite-sign}}$	0.181***	0.178***	0.146	0.428***
	(0.044)	(0.044)	(0.134)	(0.130)
Kleibergen-Paap rk F			72.958	73.479
Obs.	174238	174238	174238	174238
Panel B: Unbalanced Panel				
$\text{activism}_{\text{equal-sign}}$	0.350***	0.289***	1.000***	0.852***
	(0.038)	(0.037)	(0.087)	(0.139)
$\text{activism}_{\text{opposite-sign}}$	0.376*	0.388*	0.568*	0.765**
	(0.157)	(0.159)	(0.237)	(0.265)
Kleibergen-Paap rk F			55.895	48.649
Obs.	27652	27652	27652	27652
Controls	No	Yes	No	Yes
LegDays FE	No	Yes	No	Yes
Ind.	5808	5808	5808	5808

Note: Standard errors clustered by individuals in parenthesis. Controls include the daily average of retweets, likes, replies, and quotes. LegDays is a dummy variable that takes a value of 1 when Congress debated the abortion rights bill and 0 otherwise. Panel A: Balanced panel dataset, daily observations for one-week periods centered on legislative days. Panel B: Unbalanced panel dataset, only considering non-zero values of initial nodes' activism. * p<.05, ** p<.01, *** p<.001.

Coefficients of equal-sign activism levels are positive and significant. For instance, IV-FE estimates in Column (4) indicate that a 1-tweet increment on the equal-sign activism of peers increases initial nodes' activism by 0.38 tweets, on average. Coefficients of opposite-sign activism levels are positive and significant, except for Column (3), in which the estimate is insignificant. However, this regression corresponds to the simplest IV model without controls nor legislative days fixed effects.

[27] Except for December 29th, 2020, which window ends on January 1st, 2021.

When those are included, the estimate becomes significant. An increase of 1-tweet in the activism intensity of peers participating in the opposite online protest increases own activism by 0.43 tweets, according to Column (4).

The comparison between FE and IV-FE estimates suggests that complementarities in online activism are more substantial for the IV estimates. The difference in their magnitude is in line with the fact that these estimators compute different average treatment effects (ATE).[28] Additionally, this difference could be explained by the OLS exclusion bias and the characteristics of the compliers. Importantly, both the sign and statistical significance of the estimates remain stable among specifications.

Based on the difference in magnitude between Panel A and B estimates, one can argue that the sample restriction to non-null activism values for initial nodes leads to overestimating peer effects parameters. The coefficients in Panel B are twice as large as the analogous estimates in Panel A. In the rest of the analysis, I focus on the balanced panel dataset, where online activism includes days in which Twitter users were t-not-posting.

1.4.3 Heterogeneity analysis

This section provides two exercises to illustrate how the estimates of peer effects depend on peer groups' characteristics. Table 1.4 presents the results of the first of them: when users' exposure to early activism is taken into account. I classify as an early activist any user who posted an abortion-related tweet before the first Congress debate. The ratio of early activists at each initial node's group of peers is a source of variation in the data. I interact this ratio with peer effects parameters to see if it is relevant for understanding peer effects. Specifically, *early* is a dummy variable that takes a value of one for the individuals whose ratio of early activists in the peer group is above the sample median, 45%, and of zero otherwise.

The results suggest that the strategic complementarity between equal-sign activist peers increases as their exposure to early activism. Coefficients of the interaction between equal-sign activism and exposure are positive and significant across all specifications except Column (3). Early activism captures some degree of persistence, perhaps strength, in online activism. As such, I interpret this result as evidence of a higher complementarity between peers more involved in the abortion rights debate. In contrast, there is no evidence of a differential effect of early activism exposure in the parameters of opposite-sign activism. Accordingly, strategic complementarity between peers engaged in opposite movements does not differ based on whether the peer is a persistent activist or a newcomer. However, the interaction coefficients are not precisely estimated, as can be seen by the size of the standard errors.

[28] IV estimates the local ATE, whereas OLS estimates the ATE over the entire population.

Table 1.4 Exposure to early activism.

	FE		IV-FE	
	(1)	(2)	(3)	(4)
$\text{activism}_{\text{equal-sign}}$	0.153***	0.102***	0.528***	0.313***
	(0.016)	(0.013)	(0.028)	(0.034)
early $*$ $\text{activism}_{\text{equal-sign}}$	0.078**	0.063**	0.058	0.099**
	(0.027)	(0.021)	(0.039)	(0.038)
$\text{activism}_{\text{opposite-sign}}$	0.144***	0.133***	0.114	0.415**
	(0.032)	(0.031)	(0.144)	(0.146)
early $*$ $\text{activism}_{\text{opposite-sign}}$	0.070	0.085	0.095	0.086
	(0.085)	(0.086)	(0.263)	(0.248)
Controls	No	Yes	No	Yes
LegDays FE	No	Yes	No	Yes
Kleibergen-Paap rk F			15.878	24.637
Ind.	5808	5808	5808	5808
Obs.	174238	174238	174238	174238

Note: Standard errors clustered by individuals in parenthesis. Balanced panel dataset, daily observations for one-week periods centered on legislative days. Early is a dummy variable that takes a value of 1 if the early activist-peers ratio is above the sample median and 0 otherwise. Controls include the daily average of retweets, likes, replies, and quotes. LegDays is a dummy variable that takes a value of 1 when Congress debated the abortion rights bill and 0 otherwise. * p<.05, ** p<.01, *** p<.001.

The second exercise I perform is related to an echo chamber hypothesis. One of the main messages of Figure 1.3 is that initial nodes differ in the composition of peer groups. Around one-third of the initial nodes have no peers with a contrary viewpoint on abortion rights so we may consider them inside a chamber, i.e., belonging to a like-minded online group. On the contrary, the other two-thirds have two types of peers, with aligned and opposing viewpoints.

To consider this fact when estimating peer effects, I define *chamber* as a dummy variable that takes a value of one for the individuals whose average over time of opposite-sign peers' activism is sufficiently small and of zero otherwise. Then, by interacting this dummy variable with equal-sign activism of peers, it is possible to test the existence of an *echo* in the sub-group of initial nodes inside a *chamber*, where I interpret the existence of an echo in the lines of having a different peer effect estimate for equal-sign activism. If there is an echo effect, this interaction estimate would be higher for the sub-sample of initial nodes inside a chamber, i.e., we should observe a stronger complementarity on like-minded peers for users with no dissident peers. As seen in Table 1.5, the evidence does not support the existence of an echo-chamber phenomenon. The interaction estimate is negative but small in Columns (1)-(2), and it becomes non-statistically significant for the IV specifications.

Table 1.5 Echo chamber effect.

	FE		IV-FE	
	(1)	(2)	(3)	(4)
activism$_{\text{equal-sign}}$	0.231***	0.175***	0.549***	0.357***
	(0.017)	(0.015)	(0.033)	(0.036)
chamber $*$ activism$_{\text{equal-sign}}$	-0.072**	-0.078***	0.028	0.040
	(0.025)	(0.019)	(0.039)	(0.038)
activism$_{\text{opposite-sign}}$	0.170***	0.167***	0.162	0.449**
	(0.043)	(0.043)	(0.148)	(0.142)
Controls	No	Yes	No	Yes
LegDays FE	No	Yes	No	Yes
Kleibergen-Paap rk F			47.917	47.654
Ind.	5808	5808	5808	5808
Obs.	174238	174238	174238	174238

Note: Standard errors clustered by individuals in parenthesis. Balanced panel dataset, daily observations for one-week periods centered on legislative days. Chamber is a dummy variable that takes a value of 1 if the average over time of opposite-sign activism is <0.025 in absolute value and 0 otherwise. Controls include the daily average of retweets, likes, replies, and quotes. LegDays is a dummy variable that takes a value of 1 when Congress debated the abortion rights bill and 0 otherwise. * p<.05, ** p<.01, *** p<.001.

1.5 Robustness checks

In this section, I check the robustness of my results by relaxing the assumptions I made throughout the paper. Appendix 1.A.3 presents the corresponding results.

Unilateral links. In sections 1.2 and 1.4, I assume that links are undirected, i.e, $g_{ij} \neq 0$ if and only if $g_{ji} \neq 0$. Now, I check the sensitivity of the results to such an assumption. I perform the analysis for undirected networks - considering the peers of each initial node as the set of users who have a unilateral link with her. First, I analyze *Twitter's friends* - users followed by the initial node. Later, I consider *Twitter's followers* - users following the initial node. As seen in Appendix 1.A.3, the results remain qualitatively unchanged when considering followers as the peer group. It is true for both FE and IV-FE regressions.

Nevertheless, the results are mixed when the peer group is the set of accounts followed by the initial node - Twitter's friends. These results vary for peers with aligned and opposing viewpoints on abortion rights. In the case of like-minded peers, the results are analogous, in sign, magnitude, and statistical significance, to the ones presented in section 1.4. The estimates of opposite-sign activism of peers decrease in magnitude for the FE model and become non-statistically significant or even change their sign in the IV regressions. This result points to the importance of the proximity of peers to explain peer effects, suggesting reciprocal links and not fan-idol relationships are driving my findings.

Time span. Next, I expand the period considered in the empirical analysis. I do so to see whether the results change when the abortion rights debate becomes less salient. Specifically, I utilize the dataset with observations for a two-weeks window centered on each day Congress debated the abortion rights bill instead of one-week periods. Tables in Appendix 1.A.3 show that the results of section 1.4 are robust to the extension of the time span.

Network fixed effects. I add network fixed effects to the estimation presented in section 1.4. Specifically, I apply the local network transformation proposed by Bramoullé et al. (2009) to the activism and peers' activism variables. As can be seen in tables in Appendix 1.A.3, the main results of section 1.4 are robust to including network fixed effects.

Congress debates. Next, I split the period considered in the empirical analysis. By considering the 2018 and 2020 debates independently, it is possible to determine if peer effects and activism patterns differ between these protest periods. Tables in Appendix 1.A.3 show that activism surrounding the 2018 Congress debates leads this paper's results. There is a clear difference in data availability for one and another year, which traduces in a power loss on estimates for the 2020 debates.

1.6 Conclusion

As social media platforms have proliferated, a new public sphere where individuals connect and share ideas has emerged. Understanding how individuals engage in online interactions and how these interactions impact political outcomes is crucial for modern economies. In that regard, this paper provides novel evidence of the role of peer effects on political activism through social media platforms.

The estimates of peer effects in Section 1.4 indicate that activism exhibits strong complementarities. Remarkably, activist peers with aligned or opposing viewpoints on abortion rights have a similar effect in terms of magnitude. As mentioned, these results correspond to peer effects on the intensive margin of political activism. A natural extension of this project would also analyze the decision to be a social media activist - which posits an empirical challenge regarding its identification strategy. It will then be possible to determine whether extensive and intensive margins of activism exhibit similar patterns.

In addition, this paper suggests that the peer group's composition plays a role in understanding individual behavior - for instance, regarding exposure to early activism or the proportion of like-minded and dissident activists in the peer group. As such, social media platforms present an ideal context for further research on the influence of peers on individual behavior, as they provide detailed and precise information about social ties and online interactions. Related to this paper, some of these questions are how collective claims are created, by whom, how they evolve, and whether they persist.

1.A Appendix to Chapter 1

1.A.1 Conceptual framework

In this section, I show the assumption $|\beta| < 1$ and $|\gamma| < 1$ is a sufficient condition for the existence of $[I-\beta H+\gamma K]^{-1}$, which allows me to write (1.2.4). The proof consists of two steps. First, demonstrate that provided $|\beta| < 1$ and $|\gamma| < 1$, the matrix $[I-\beta H+\gamma K]$ is a *strictly diagonally dominant* matrix. Then, apply the *Gershgorin's circle theorem* to argue that the matrix is non-singular and, consequently, that its inverse exists.

A square matrix is said to be strictly diagonally dominant if, for every row, its diagonal entry is larger than the sum of the absolute values of the non-diagonal entries in that row. That is, A is strictly diagonally dominant if

$$|a_{ii}| > \sum_{j\neq i} |a_{ij}| \quad \forall i$$

The diagonal entries of $[I - \beta H + \gamma K]$ are equal to 1, whereas the non-diagonal entries are either $-\beta h_{ij}$ or γk_{ij}. So, this matrix is strictly diagonally dominant if

$$\begin{aligned} 1 >& \sum_{j\neq i} |\beta h_{ij}| + \sum_{j\neq i} |\gamma k_{ij}| \quad \forall i \\ 1 >& |\beta| \sum_{j\neq i} h_{ij} + |\gamma| \sum_{j\neq i} k_{ij} \quad \forall i \end{aligned}$$

Where the second step follows from properties of absolute value and the fact that the entries of H and K are non-negative. Furthermore, as $G = H + K$, and G is row-normalized, it holds that

$$\sum_j h_{ij} + \sum_j k_{ij} = \sum_j g_{ij} = 1 \quad \forall i$$

Then, the right-hand side of the above inequality is a linear combination of $|\beta|$ and $|\gamma|$, and the condition of $|\beta| < 1$ and $|\gamma| < 1$ is sufficient to guarantee the inequality holds. It follows that $[I - \beta H + \gamma K]$ is strictly diagonally dominant, and that $[I - \beta H + \gamma K]^{-1}$ exists. As the inverse is unique, a unique vector $\mathbf{a}$ is compatible with equation (1.2.4).

1.A.2 Twitter data

Twitter data collection

Twitter is an online platform that allows users to publish short messages, of a maximum of 140 characters, on their profiles. In January 2021, Twitter launched an Academic Research product track, which enables researchers to access all v2 endpoints. Notably, the *Twitter Search API v2* gives access to the entire history of public conversations and not only recent tweets. For more information about the academic track on Twitter, follow this link. I collected Twitter data with the command line tool and Python library, twarc2.

Tweets collection To collect tweets, I relied on the *v2 full-archive search endpoint*. I constructed the Twitter query to include all the tweets in Spanish, net of retweets, which include at least one of the hashtags present in Table 1.6.

Table 1.6 List of hashtags considered in the Twitter query.

Pro-choice hashtags	Pro-life hashtags
#AbortoLegalYa	#ArgentinaEsProvida
#AbortoLegal	#ArgentinaProVida
#AbortoLegalSeguroyGratuito	#AbortoCero
#AbortoLegalYSeguro	#DefendamosLaVida
#AbortoLibre	#LegaloIlegalelAbortoMataIgual
#AbortoVoluntario	#MarchaPorLaVida
#AbortarEnPandemia*	#NoAlAborto
#EsLey*	#OlaCeleste
#GarantizarDerechosNoEsDelito	#PañueloCeleste
#IVE	#SalvemosLasDosVidas
#LaOlaVerde	#SalvemosLas2Vidas
#MareaVerde	#SalvenALos2
#PañueloVerde	#SiALaVida
#QueSubaLaMarea	#SoyProvida
#SeraLey	#TodaVidaVale
#UnaConquistaFeminista*	

Collection date: September 2021. *For 2020 only.

User data collection To collect Twitter data relative to users, I relied on the *follows lookup endpoints*. For any user of interest, I requested the list of her friends (following) and followers. To obtain mutual connections, I intersected these lists.

1.A.3 Tables and Figures

Figure 1.5 Pro-choice and pro-life handkerchiefs.

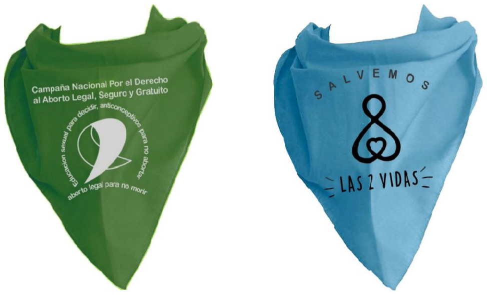

Campaña Nacional Por el Derecho al Aborto Legal, Seguro y Gratuito
Educación sexual para decidir, anticonceptivos para no abortar
aborto legal para no morir
SALVEMOS
LAS 2 VIDAS

Figure 1.6 Correlation between activism and the activist-peers ratio.

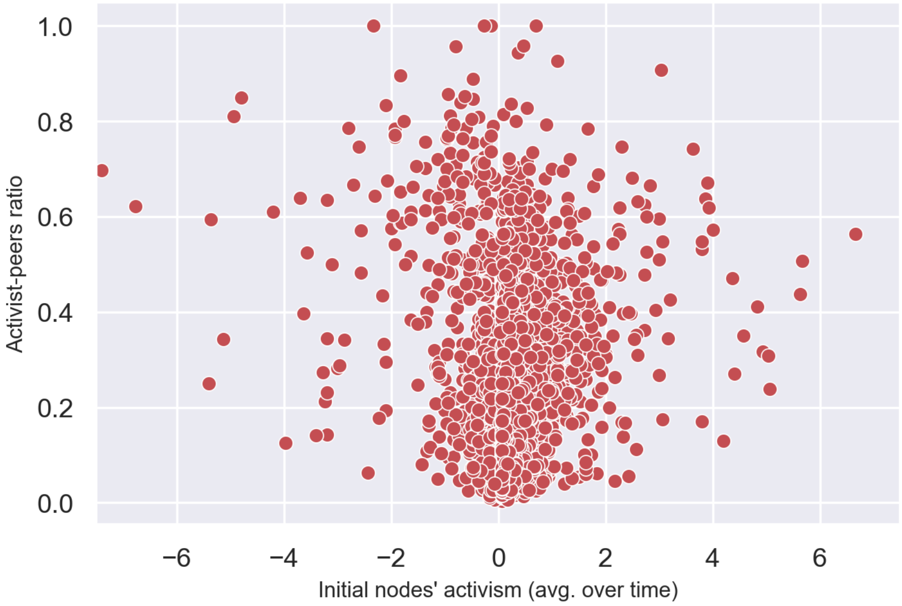


Figure 1.7 Activism histograms. Initial nodes and their peers.

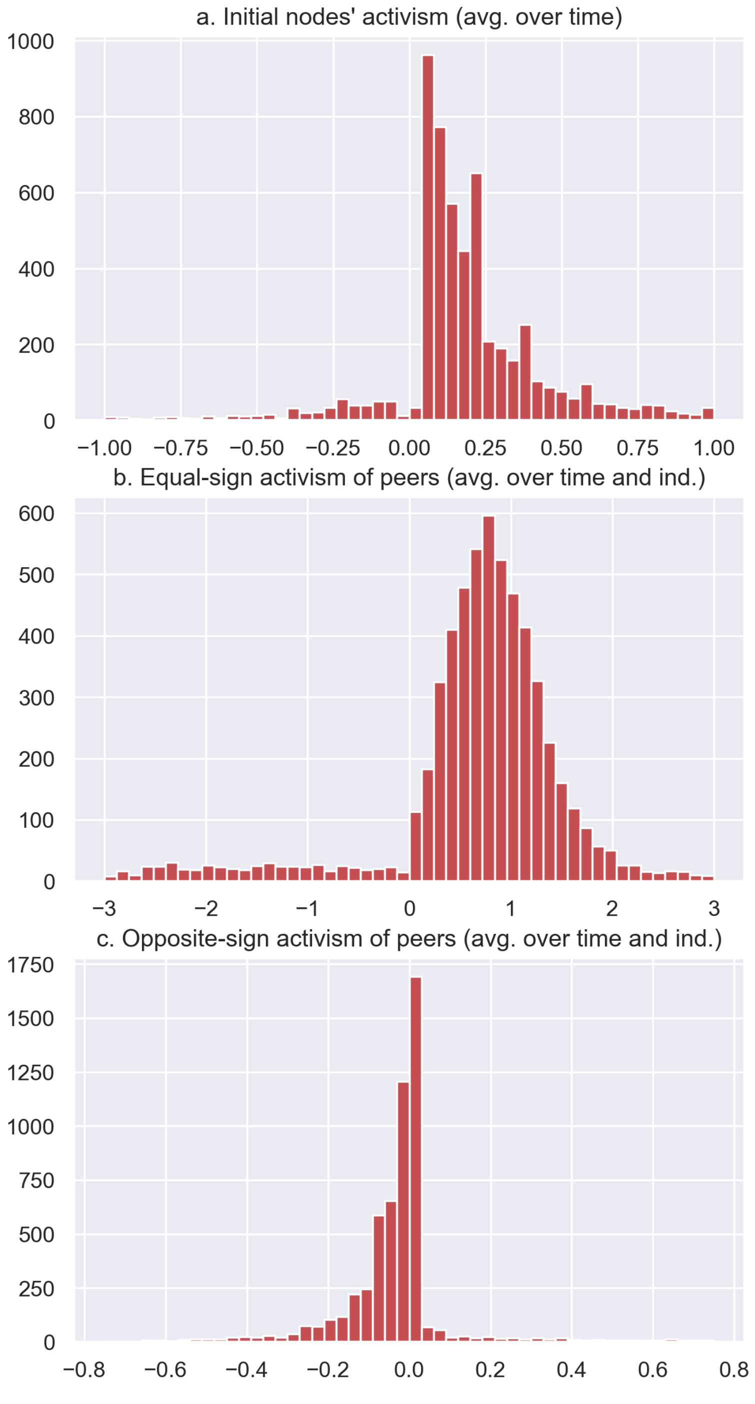


Figure 1.8 Activist peers histograms. Pro-choice initial nodes.

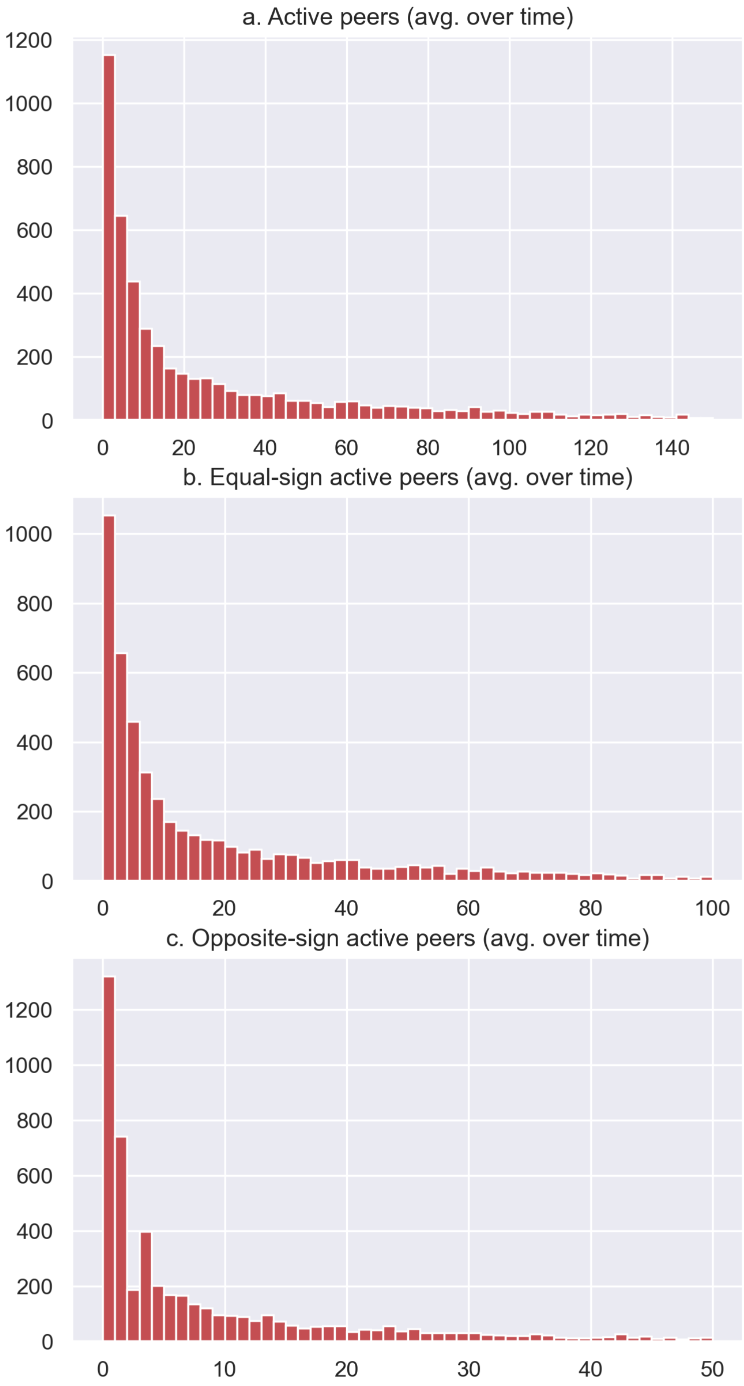
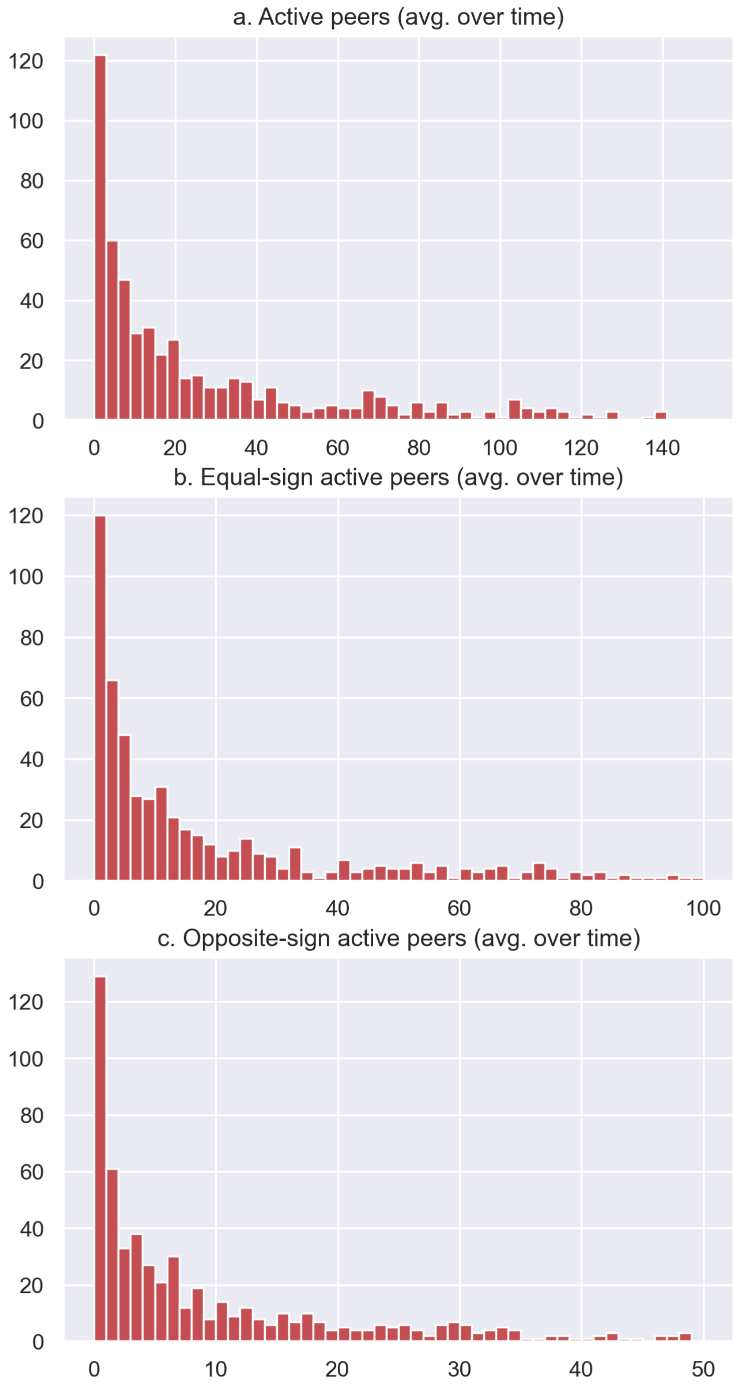


Figure 1.9 Activist peers histograms. Pro-life initial nodes.



Increasing the time span

Table 1.7 Peer effects in online activism. Two-weeks period.

	FE		IV-FE	
	(1)	(2)	(3)	(4)
Panel A: Balanced Panel				
$\text{activism}_{\text{equal-sign}}$	0.192***	0.138***	0.508***	0.360***
	(0.011)	(0.010)	(0.016)	(0.022)
$\text{activism}_{\text{opposite-sign}}$	0.165***	0.161***	0.302**	0.482***
	(0.095)	(0.098)	(0.194)	(0.219)
Kleibergen-Paap rk F			71.263	70.929
Obs.	354345	354345	354345	354345
Panel B: Unbalanced Panel				
$\text{activism}_{\text{equal-sign}}$	0.408***	0.350***	0.936***	0.799***
	(0.041)	(0.041)	(0.070)	(0.108)
$\text{activism}_{\text{opposite-sign}}$	0.301**	0.314**	0.719***	0.910***
	(0.095)	(0.098)	(0.194)	(0.219)
Kleibergen-Paap rk F			66.815	61.305
Obs.	33597	33597	33597	33597
Controls	No	Yes	No	Yes
LegDays FE	No	Yes	No	Yes
Ind.	5809	5809	5809	5809

Note: Standard errors clustered by individuals in parenthesis. Panel A: Balanced panel dataset, daily observations for two-week periods centered on legislative days. Panel B: Unbalanced panel dataset, only considering non-zero values of initial nodes' activism. Controls include the daily average of retweets, likes, replies, and quotes. LegDays is a dummy variable that takes value 1 when Congress debated the abortion rights bill and 0 otherwise. * p<.05, ** p<.01, *** p<.001.

Table 1.8 Exposure to early activism. Two-weeks period.

	FE		IV-FE	
	(1)	(2)	(3)	(4)
$\text{activism}_{\text{equal-sign}}$	0.155***	0.107***	0.470***	0.300***
	(0.013)	(0.011)	(0.025)	(0.032)
$\text{early} * \text{activism}_{\text{equal-sign}}$	0.071***	0.059***	0.062	0.094**
	(0.021)	(0.017)	(0.033)	(0.033)
$\text{activism}_{\text{opposite-sign}}$	0.159***	0.147***	0.253*	0.454***
	(0.031)	(0.030)	(0.119)	(0.121)
$\text{early} * \text{activism}_{\text{opposite-sign}}$	0.013	0.027	0.132	0.118
	(0.050)	(0.050)	(0.186)	(0.182)
Controls	No	Yes	No	Yes
LegDays FE	No	Yes	No	Yes
Kleibergen-Paap rk F			13.947	14.864
Ind.	5809	5809	5809	5809
Obs.	354345	354345	354345	354345

Note: Standard errors clustered by individuals in parenthesis. Balanced panel dataset, daily observations for two-week periods centered on legislative days. Early is a dummy variable that takes a value of 1 if the early activist-peers ratio is above the sample median and 0 otherwise. Controls include the daily average of retweets, likes, replies, and quotes. LegDays is a dummy variable that takes a value of 1 when Congress debated the abortion rights bill and 0 otherwise. * p<.05, ** p<.01, *** p<.001.

Table 1.9 Echo chamber effect. Two-weeks period.

	FE		IV-FE	
	(1)	(2)	(3)	(4)
$\text{activism}_{\text{equal-sign}}$	0.196***	0.156***	0.443***	0.295***
	(0.014)	(0.013)	(0.032)	(0.037)
$\text{chamber} * \text{activism}_{\text{equal-sign}}$	-0.008	-0.034*	0.103**	0.106**
	(0.021)	(0.017)	(0.036)	(0.036)
$\text{activism}_{\text{opposite-sign}}$	0.164***	0.154***	0.374***	0.554***
	(0.025)	(0.024)	(0.111)	(0.112)
Controls	No	Yes	No	Yes
LegDays FE	No	Yes	No	Yes
Kleibergen-Paap rk F			44.921	44.021
Ind.	5809	5809	5809	5809
Obs.	354345	354345	354345	354345

Note: Standard errors clustered by individuals in parenthesis. Balanced panel dataset, daily observations for two-week periods centered on legislative days. Chamber is a dummy variable that takes a value of 1 if the average over time of opposite-sign activism is <0.025, in absolute value, and 0 otherwise. Controls include the daily average of retweets, likes, replies, and quotes. LegDays is a dummy variable that takes a value of 1 when Congress debated the abortion rights bill and 0 otherwise. * p<.05, ** p<.01, *** p<.001.

Considering unilateral links

Table 1.10 Peer effects in online activism. Friends as peers.

	FE		IV-FE	
	(1)	(2)	(3)	(4)
Panel A: Balanced Panel				
$\text{activism}_{\text{equal-sign}}$	0.152***	0.101***	0.480***	0.311***
	(0.006)	(0.006)	(0.028)	(0.029)
$\text{activism}_{\text{opposite-sign}}$	0.094***	0.076***	-0.158	0.107
	(0.011)	(0.011)	(0.103)	(0.095)
Kleibergen-Paap rk F			50.770	55.028
Obs.	173998	173998	173998	173998
Panel B: Unbalanced Panel				
$\text{activism}_{\text{equal-sign}}$	0.337***	0.284***	0.990***	0.940***
	(0.027)	(0.027)	(0.134)	(0.170)
$\text{activism}_{\text{opposite-sign}}$	0.134***	0.122***	0.017	0.088
	(0.024)	(0.025)	(0.230)	(0.238)
Kleibergen-Paap rk F			24.811	22.684
Obs.	27607	27607	27607	27607
Controls	No	Yes	No	Yes
LegDays FE	No	Yes	No	Yes
Ind.	5800	5800	5800	5800

Note: Standard errors clustered by individuals in parenthesis. Panel A: Balanced panel dataset, daily observations for one-week periods centered on legislative days. Panel B: Unbalanced panel dataset, only considering non-zero values of initial nodes' activism. The peer group is the set of Twitter accounts followed by the individual. Controls include the daily average of retweets, likes, replies, and quotes. LegDays is a dummy variable that takes a value of 1 when Congress debated the abortion rights bill and 0 otherwise. * p<.05, ** p<.01, *** p<.001.

Table 1.11 Peer effects in online activism. Followers as peers.

	FE		IV-FE	
	(1)	(2)	(3)	(4)
Panel A: Balanced Panel				
$\text{activism}_{\text{equal-sign}}$	0.211***	0.148***	0.568***	0.373***
	(0.014)	(0.012)	(0.022)	(0.027)
$\text{activism}_{\text{opposite-sign}}$	0.115***	0.103***	0.097	0.365**
	(0.017)	(0.016)	(0.115)	(0.112)
Kleibergen-Paap rk F			69.105	70.328
Obs.	174119	174119	174119	174119
Panel B: Unbalanced Panel				
$\text{activism}_{\text{equal-sign}}$	0.412***	0.342***	0.997***	0.872***
	(0.038)	(0.038)	(0.091)	(0.148)
$\text{activism}_{\text{opposite-sign}}$	0.172***	0.176***	0.560**	0.711**
	(0.052)	(0.053)	(0.211)	(0.237)
Kleibergen-Paap rk F			58.843	54.466
Obs.	27632	27632	27632	27632
Controls	No	Yes	No	Yes
LegDays FE	No	Yes	No	Yes
Ind.	5804	5804	5804	5804

Note: Standard errors clustered by individuals in parenthesis. Panel A: Balanced panel dataset, daily observations for one-week periods centered on legislative days. Panel B: Unbalanced panel dataset, only considering non-zero values of initial nodes' activism. The peer group is the set of Twitter accounts that follow the individual. Controls include the daily average of retweets, likes, replies, and quotes. LegDays is a dummy variable that takes a value of 1 when Congress debated the abortion rights bill and 0 otherwise. * p<.05, ** p<.01, *** p<.001.

Table 1.12 Exposure to early activism. Unilateral links.

	FE		IV-FE	
	(1)	(2)	(3)	(4)
Panel A: Friends as peers				
$\text{activism}_{\text{equal-sign}}$	0.086***	0.057***	0.470***	0.290***
	(0.009)	(0.008)	(0.043)	(0.048)
$\text{early} * \text{activism}_{\text{equal-sign}}$	0.089***	0.060***	0.014	0.025
	(0.012)	(0.010)	(0.054)	(0.054)
$\text{activism}_{\text{opposite-sign}}$	0.148***	0.138***	-0.102	0.185
	(0.027)	(0.026)	(0.159)	(0.158)
$\text{early} * \text{activism}_{\text{opposite-sign}}$	-0.063*	-0.070*	-0.068	-0.091
	(0.029)	(0.028)	(0.198)	(0.189)
Kleibergen-Paap rk F			19.516	21.812
Ind.	5800	5800	5800	5800
Obs.	173998	173998	173998	173998
Panel B: Followers as peers				
$\text{activism}_{\text{equal-sign}}$	0.177***	0.124***	0.558***	0.352***
	(0.016)	(0.013)	(0.030)	(0.035)
$\text{early} * \text{activism}_{\text{equal-sign}}$	0.081**	0.058*	0.018	0.039
	(0.029)	(0.022)	(0.045)	(0.044)
$\text{activism}_{\text{opposite-sign}}$	0.125***	0.110***	0.088	0.349**
	(0.023)	(0.021)	(0.124)	(0.115)
$\text{early} * \text{activism}_{\text{opposite-sign}}$	-0.025	-0.017	0.025	0.048
	(0.035)	(0.032)	(0.244)	(0.237)
Kleibergen-Paap rk F			18.329	18.811
Ind.	5804	5804	5804	5804
Obs.	174119	174119	174119	174119
Controls	No	Yes	No	Yes
LegDays FE	No	Yes	No	Yes

Note: Standard errors clustered by individuals in parenthesis. Balanced panel dataset. Daily observations for one-week periods centered on legislative days. Panel A: The peer group is the set of Twitter accounts followed by the individual. Panel B: The peer group is the set of Twitter accounts that follow the individual. Early is a dummy variable that takes a value of 1 if the early activist-peers ratio is above the sample median and 0 otherwise. Controls include the daily average of retweets, likes, replies, and quotes. LegDays is a dummy variable that takes value 1 when Congress debated the abortion rights bill and 0 otherwise. * p<.05, ** p<.01, *** p<.001.

Table 1.13 Echo chamber effect. Unilateral links.

	FE		IV-FE	
	(1)	(2)	(3)	(4)
Panel A: Friends as peers				
$\text{activism}_{\text{equal-sign}}$	0.163***	0.113***	0.507***	0.320***
	(0.007)	(0.007)	(0.036)	(0.035)
$\text{chamber} * \text{activism}_{\text{equal-sign}}$	-0.049***	-0.051***	-0.109**	-0.043
	(0.014)	(0.011)	(0.041)	(0.040)
$\text{activism}_{\text{opposite-sign}}$	0.091***	0.073***	-0.187	0.098
	(0.011)	(0.011)	(0.112)	(0.101)
Kleibergen-Paap rk F			30.677	33.832
Ind.	5800	5800	5800	5800
Obs.	173998	173998	173998	173998
Panel B: Followers as peers				
$\text{activism}_{\text{equal-sign}}$	0.222***	0.164***	0.563***	0.361***
	(0.018)	(0.015)	(0.029)	(0.033)
$\text{chamber} * \text{activism}_{\text{equal-sign}}$	-0.034	-0.054**	0.013	0.036
	(0.027)	(0.020)	(0.042)	(0.040)
$\text{activism}_{\text{opposite-sign}}$	0.112***	0.098***	0.101	0.374**
	(0.017)	(0.016)	(0.121)	(0.117)
Kleibergen-Paap rk F			43.874	44.400
Ind.	5804	5804	5804	5804
Obs.	174119	174119	174119	174119
Controls	No	Yes	No	Yes
LegDays FE	No	Yes	No	Yes

Note: Standard errors clustered by individuals in parenthesis. Balanced panel dataset, daily observations for one-week periods centered on legislative days. Panel A: The peer group is the set of Twitter accounts followed by the individual. Panel B: The peer group is the set of Twitter accounts that follow the individual. Chamber is a dummy variable that takes a value of 1 if the average over time of opposite-sign activism is <0.025 in absolute value and 0 otherwise. Controls include the daily average of retweets, likes, replies, and quotes. LegDays is a dummy variable that takes value 1 when Congress debated the abortion rights bill and 0 otherwise. * p<.05, ** p<.01, *** p<.001.

Adding network fixed effects

Table 1.14 Peer effects in online activism. Network FE.

	FE		IV-FE	
	(1)	(2)	(3)	(4)
Panel A: Balanced Panel				
$\text{activism}_{\text{equal-sign}}$	0.125***	0.069***	0.548***	0.288***
	(0.011)	(0.008)	(0.035)	(0.042)
$\text{activism}_{\text{opposite-sign}}$	0.259***	0.244***	0.218	0.584***
	(0.048)	(0.047)	(0.148)	(0.145)
Kleibergen-Paap rk F			68.047	65.852
Obs.	174238	174238	174238	174238
Panel B: Unbalanced Panel				
$\text{activism}_{\text{equal-sign}}$	0.135***	0.079*	0.832***	0.353
	(0.035)	(0.033)	(0.185)	(0.341)
$\text{activism}_{\text{opposite-sign}}$	0.456**	0.453**	0.751*	1.256**
	(0.148)	(0.148)	(0.315)	(0.431)
Kleibergen-Paap rk F			39.540	14.895
Obs.	27652	27652	27652	27652
Controls	No	Yes	No	Yes
LegDays FE	No	Yes	No	Yes
Network FE	Yes	Yes	Yes	Yes
Ind.	5808	5808	5808	5808

Note: Standard errors clustered by individuals in parenthesis. Panel A: Balanced panel dataset, daily observations for one-week periods centered on legislative days. Panel B: Unbalanced panel dataset, only considering non-zero values of initial nodes' activism. Controls include the daily average of retweets, likes, replies, and quotes. LegDays is a dummy variable that takes a value of 1 when Congress debated the abortion rights bill and 0 otherwise. * p<.05, ** p<.01, *** p<.001.

Table 1.15 Exposure to early activism. Network FE.

	FE		IV-FE	
	(1)	(2)	(3)	(4)
activism$_{\text{equal-sign}}$	0.111***	0.060***	0.521***	0.260***
	(0.014)	(0.010)	(0.038)	(0.046)
early $*$ activism$_{\text{equal-sign}}$	0.028	0.018	0.039	0.036
	(0.022)	(0.015)	(0.069)	(0.064)
activism$_{\text{opposite-sign}}$	0.211***	0.202***	0.158	0.508**
	(0.044)	(0.044)	(0.153)	(0.155)
early $*$ activism$_{\text{opposite-sign}}$	0.086	0.077	0.148	0.185
	(0.093)	(0.091)	(0.291)	(0.264)
Controls	No	Yes	No	Yes
LegDays FE	No	Yes	No	Yes
Network FE	Yes	Yes	Yes	Yes
Kleibergen-Paap rk F			17.689	23.419
Ind.	5808	5808	5808	5808
Obs.	174238	174238	174238	174238

Note: Standard errors clustered by individuals in parenthesis. Balanced panel dataset. Daily observations for one-week periods centered on legislative days. Early is a dummy variable that takes a value of 1 if the early activist-peers ratio is above the sample median and 0 otherwise. Controls include the daily average of retweets, likes, replies, and quotes. LegDays is a dummy variable that takes value 1 when Congress debated the abortion rights bill and 0 otherwise. * p<.05, ** p<.01, *** p<.001.

Table 1.16 Echo chamber effect. Network FE.

	FE		IV-FE	
	(1)	(2)	(3)	(4)
$\text{activism}_{\text{equal-sign}}$	0.156***	0.102***	0.514***	0.230***
	(0.014)	(0.011)	(0.060)	(0.064)
$\text{chamber} * \text{activism}_{\text{equal-sign}}$	-0.060**	-0.065***	0.060	0.106
	(0.020)	(0.014)	(0.057)	(0.054)
$\text{activism}_{\text{opposite-sign}}$	0.249***	0.233***	0.253	0.644***
	(0.048)	(0.047)	(0.174)	(0.170)
Controls	No	Yes	No	Yes
LegDays FE	No	Yes	No	Yes
Network FE	Yes	Yes	Yes	Yes
Kleibergen-Paap rk F			42.940	40.970
Ind.	5808	5808	5808	5808
Obs.	174238	174238	174238	174238

Note: Standard errors clustered by individuals in parenthesis. Balanced panel dataset, daily observations for one-week periods centered on legislative days. Chamber is a dummy variable that takes a value of 1 if the average over time of opposite-sign activism is <0.025, in absolute value, and 0 otherwise. Controls include the daily average of retweets, likes, replies, and quotes. LegDays is a dummy variable that takes value 1 when Congress debated the abortion rights bill and 0 otherwise. * p<.05, ** p<.01, *** p<.001.

Congress debates

Table 1.17 Peer effects in online activism. 2018 Congress debates.

	FE		IV-FE	
	(1)	(2)	(3)	(4)
Panel A: Balanced Panel				
$\text{activism}_{\text{equal-sign}}$	0.231***	0.151***	0.666***	0.404***
	(0.017)	(0.014)	(0.025)	(0.035)
$\text{activism}_{\text{opposite-sign}}$	0.208**	0.206**	0.276*	0.619***
	(0.067)	(0.067)	(0.138)	(0.139)
Kleibergen-Paap rk F			62.334	62.618
Ind.	5808	5808	5808	5808
Obs.	92927	92927	92927	92927
Panel B: Unbalanced Panel				
$\text{activism}_{\text{equal-sign}}$	0.506***	0.399***	1.322***	1.155***
	(0.066)	(0.063)	(0.107)	(0.215)
$\text{activism}_{\text{opposite-sign}}$	0.409	0.434	0.432	0.644
	(0.240)	(0.248)	(0.272)	(0.332)
Kleibergen-Paap rk F			28.364	23.587
Ind.	3912	3912	3912	3912
Obs.	15464	15464	15464	15464
Controls	No	Yes	No	Yes
LegDays FE	No	Yes	No	Yes

Note: Standard errors clustered by individuals in parenthesis. Panel A: Balanced panel dataset, daily observations for one-week periods centered on legislative days. Panel B: Unbalanced panel dataset, only considering non-zero values of initial nodes' activism. Controls include the daily average of retweets, likes, replies, and quotes. LegDays is a dummy variable that takes a value of 1 when Congress debated the abortion rights bill and 0 otherwise. * p<.05, ** p<.01, *** p<.001.

Table 1.18 Peer effects in online activism. 2020 Congress debates.

	FE		IV-FE	
	(1)	(2)	(3)	(4)
Panel A: Balanced Panel				
$\text{activism}_{\text{equal-sign}}$	0.163***	0.112***	0.466***	0.340***
	(0.014)	(0.012)	(0.020)	(0.025)
$\text{activism}_{\text{opposite-sign}}$	0.137***	0.142***	-0.100	0.140
	(0.038)	(0.039)	(0.172)	(0.176)
Kleibergen-Paap rk F			45.611	44.907
Ind.	5808	5808	5808	5808
Obs.	81311	81311	81311	81311
Panel B: Unbalanced Panel				
$\text{activism}_{\text{equal-sign}}$	0.371***	0.345***	0.685***	0.572***
	(0.066)	(0.068)	(0.137)	(0.142)
$\text{activism}_{\text{opposite-sign}}$	0.464**	0.488**	1.457*	1.700**
	(0.166)	(0.168)	(0.582)	(0.612)
Kleibergen-Paap rk F			18.668	17.114
Ind.	2432	2432	2432	2432
Obs.	6924	6924	6924	6924
Controls	No	Yes	No	Yes
LegDays FE	No	Yes	No	Yes

Note: Standard errors clustered by individuals in parenthesis. Panel A: Balanced panel dataset, daily observations for one-week periods centered on legislative days. Panel B: Unbalanced panel dataset, only considering non-zero values of initial nodes' activism. Controls include the daily average of retweets, likes, replies, and quotes. LegDays is a dummy variable that takes a value of 1 when Congress debated the abortion rights bill and 0 otherwise. * p<.05, ** p<.01, *** p<.001.

Table 1.19 Exposure to early activism. Congress debates.

	FE		IV-FE	
	(1)	(2)	(3)	(4)
Panel A: 2018				
$\text{activism}_{\text{equal-sign}}$	0.177***	0.108***	0.596***	0.290***
	(0.018)	(0.015)	(0.036)	(0.046)
$\text{early} * \text{activism}_{\text{equal-sign}}$	0.103**	0.084**	0.112*	0.177***
	(0.035)	(0.027)	(0.048)	(0.048)
$\text{activism}_{\text{opposite-sign}}$	0.179***	0.162***	0.353	0.753***
	(0.037)	(0.035)	(0.188)	(0.186)
$\text{early} * \text{activism}_{\text{opposite-sign}}$	0.042	0.066	-0.093	-0.161
	(0.110)	(0.110)	(0.271)	(0.266)
Kleibergen-Paap rk F			17.367	18.684
Obs.	92927	92927	92927	92927
Panel B: 2020				
$\text{activism}_{\text{equal-sign}}$	0.144***	0.101***	0.448***	0.329***
	(0.019)	(0.015)	(0.030)	(0.035)
$\text{early} * \text{activism}_{\text{equal-sign}}$	0.038	0.021	0.033	0.021
	(0.028)	(0.021)	(0.040)	(0.039)
$\text{activism}_{\text{opposite-sign}}$	0.113**	0.114**	-0.025	0.152
	(0.043)	(0.044)	(0.224)	(0.229)
$\text{early} * \text{activism}_{\text{opposite-sign}}$	0.083	0.093	-0.16	-0.021
	(0.058)	(0.059)	(0.353)	(0.343)
Kleibergen-Paap rk F			10.658	10.967
Obs.	81311	81311	81311	81311
Controls	No	Yes	No	Yes
LegDays FE	No	Yes	No	Yes
Ind.	5808	5808	5808	5808

Note: Standard errors clustered by individuals in parenthesis. Balanced panel dataset. Daily observations for one-week periods centered on legislative days. Early is a dummy variable that takes a value of 1 if the early activist-peers ratio is above the sample median and 0 otherwise. Controls include the daily average of retweets, likes, replies, and quotes. LegDays is a dummy variable that takes value 1 when Congress debated the abortion rights bill and 0 otherwise. * p<.05, ** p<.01, *** p<.001.

Table 1.20 Echo chamber effect. Congress debates.

	FE		IV-FE	
	(1)	(2)	(3)	(4)
Panel A: 2018				
$\text{activism}_{\text{equal-sign}}$	0.247***	0.180***	0.661***	0.384***
	(0.027)	(0.023)	(0.040)	(0.049)
$\text{chamber} * \text{activism}_{\text{equal-sign}}$	(0.037)	-0.064*	0.008	0.039
	(0.033)	(0.025)	(0.049)	(0.048)
$\text{activism}_{\text{opposite-sign}}$	0.203**	0.198**	0.280	0.635***
	(0.067)	(0.067)	(0.149)	(0.149)
Kleibergen-Paap rk F			42.130	41.488
Obs.	92927	92927	92927	92927
Panel B: 2020				
$\text{activism}_{\text{equal-sign}}$	0.210***	0.157***	0.468***	0.345***
	(0.013)	(0.013)	(0.029)	(0.033)
$\text{chamber} * \text{activism}_{\text{equal-sign}}$	-0.087***	-0.081***	-0.003	-0.010
	(0.023)	(0.017)	(0.038)	(0.037)
$\text{activism}_{\text{opposite-sign}}$	0.123***	0.128***	-0.102	0.134
	(0.034)	(0.036)	(0.185)	(0.188)
Kleibergen-Paap rk F			29.348	28.784
Obs.	81311	81311	81311	81311
Controls	No	Yes	No	Yes
LegDays FE	No	Yes	No	Yes
Ind.	5808	5808	5808	5808

Note: Standard errors clustered by individuals in parenthesis. Balanced panel dataset, daily observations for one-week periods centered on legislative days. Chamber is a dummy variable that takes a value of 1 if the average over time of opposite-sign activism is <0.025 in absolute value and 0 otherwise. Controls include the daily average of retweets, likes, replies, and quotes. LegDays is a dummy variable that takes value 1 when Congress debated the abortion rights bill and 0 otherwise. * p<.05, ** p<.01, *** p<.001.

Chapter 2

Discrete Social Norms in Networks

2.1 Introduction

Social norms encompass unwritten rules and informal agreements, vital in governing our interactions with others. In this context, others may refer to group members, such as friends, colleagues, neighbors, or the entire society, since social norms permeate almost every aspect of our lives. These norms are sustained through diverse mechanisms, including the necessity of coordination, the fear of ostracism, and the demonstration of team membership, among others. This paper focuses on a particular mechanism, *local conformity* - as individuals conform to the social norm of their peer groups. Accordingly, social norms refer to behavior patterns that group members uphold and reinforce: they conform to the norm and simultaneously expect everyone else to conform.

In this paper, I assume group interactions are structured through a social network. Within this framework, I develop a model of conformity and social norms. In the model, social norms are *local* and *endogenous.* They are local as each individual considers her social norm to be the *average action of her peers.* They are endogenous as long as each individual takes her peers as a reference group when choosing how to behave, so social norms are simultaneously determined in equilibrium.

Social norms establish standards on different aspects of life, ranging from dress code and contractual relationships to conceptions of right and wrong, fairness, and equality. Some of these codes of conduct are, by nature, discrete. Regarding female labor market participation, a woman either works full or part-time or does not do it; individuals are either in favor or against public policies such as abortion access or same-sex marriage; a traditional example of harmful social norm relates to female genital cutting; regarding religion, individuals choose whether to have one or not (and which one).

Nonetheless, the economic literature on conformity and social networks mainly has analyzed continuous actions - Patacchini and Zenou (2012), Boucher (2016), and Ushchev and Zenou (2020). In this paper, I focus on the case where actions are *discrete*, shedding light on such scenarios and filling a gap in the literature. I show that, under the assumption of a discrete choice set, the social norms game generally admits a multiplicity of equilibria. This result contrasts the model with continuous actions, characterized by the uniqueness of equilibrium.

Remarkably, the interplay between individual preferences, the network structure, and the taste for conformity leads to three types of equilibrium. A purely conformist equilibrium is such that all individuals choose the same code of conduct. In a pure individualist equilibrium, agents choose actions based solely on their preferences. Finally, a partially conformist-individualist equilibrium is characterized by a subset of individuals conforming to social norms while others choose actions based on their preferences.

The model reveals that the multiplicity of equilibria arises when society strongly favors conformity, implying that different social norms might materialize. In such cases, in despite of the local configuration of the game, global social norms may emerge, as purely conformist equilibrium exists. On the contrary, a unique equilibrium exists if society's taste for conformity is relatively low. This unique equilibrium exhibits strong levels of individualism, but some individuals conform to their local social norms under specific network structures.

Related literature. This paper contributes to the growing literature on conformity in network games, mainly based on the local-average model. Patacchini and Zenou (2012) prove the existence and uniqueness of the Nash equilibrium for this model and use it to empirically study how conformism affects juvenile crime. Boucher (2016) theoretically and empirically studies a network formation model with conformity. Ushchev and Zenou (2020) deeply discuss the properties, comparative statics, and welfare implications of the local-average model. Genicot (2022) proposes a slightly different model to study how individuals' tolerance affects their ability to compromise. Unlike the mentioned papers, I assume individuals make discrete decisions, given the (exogenous) network in which they are embedded. Then, I concentrate my analysis on understanding how different assumptions on the preferences space and choice set impact the equilibrium characterization of the game. More broadly, this paper speaks to the literature on conformity and social norms (for example, Bernheim (1994), Young (2015), and Gulesci et al. (2021)), and to the literature on network games,[1] especially those under strategic complementarity (Ballester and Calvó-Armengol (2010) and Belhaj et al. (2014), among others).

The rest of the paper is organized as follows. Section 2.2 introduces model preliminaries. Section 2.3 and 2.4 present the theoretical model when social norms

[1] See Jackson and Zenou (2015) for a review.

are continuous (benchmark) and discrete, respectively. Section 2.5 concludes.

2.2 Model preliminaries

Consider a society comprised of $n < \infty$ individuals who are embedded in a network g. Let $G = [g_{ij}]$ be an $n \times n$ non-negative adjacency matrix representing links between individuals. The (i, j) entry of G, denoted g_{ij}, equals 1 if individuals i and j have a link and 0 otherwise. Each individual i has a specific peer group of size $d_i = \sum_{j=1}^{n} g_{ij}$, where I refer d_i as individual i's degree. I assume the network g is *undirected*, i.e., $g_{ij} = g_{ji} \quad \forall(i, j)$, and has *no self-loops*, i.e., $g_{ii} = 0 \quad \forall i$. Define $\hat{G} = [\hat{g}_{ij}]$ as the $n \times n$ row-normalized adjacency matrix, such that $g_{ij} = 1/d_i$ if individuals i and j have a link and $\hat{g}_{ij} = 0$ otherwise.

In addition to her peer group, each individual i is characterized by an exogenous preference parameter, denoted $\alpha_i \in \mathbb{R}_+$. Conditional on the network structure and her preferences, individual i chooses an action, denoted by x_i. Throughout this paper, I will make different assumptions about *choice set* - particularly regarding its discreteness or continuity. Finally, the (local) social norm of individual i, denoted by $\overline{x}_i$, is the average action across her peers, namely,

$$\overline{x}_i \equiv \sum_{j=1}^{n} \hat{g}_{ij} x_j$$

Following the literature on conformism in network games, e.g., Patacchini and Zenou (2012), Boucher (2016), and Ushchev and Zenou (2020), I assume a linear quadratic specification for the utility function. Denoting any profile of action by $\mathbf{x}$, the following function represents i's utility:

$$U_i\left(x_i, \mathbf{x}_{-i}, \mathbf{g}\right) = \alpha_i x_i - \frac{1}{2} x_i^2 - \frac{\theta}{2}\left(x_i - \bar{x}_i\right)^2 \tag{2.2.1}$$

The first two terms of equation (2.2.1) reflect i's private benefit and cost associated with her chosen action. The third term represents the social cost of choosing an action different from i's social norm, i.e., the average action of her peers. The parameter $\theta > 0$ measures the *taste for conformity* in society. Equation (2.2.1) is usually named as the *local-average model*, as individuals want to conform to their local-average action. Note that, under this specification, the effect of conformism on an individual utility is independent of the number of peers she has.[2]

Individuals play a non-cooperative game, conditional on the network structure

[2] Boucher (2016) proposes a different specification for the utility function, which accounts for the dependency of conformism on the cardinality of peer groups. In that case, individuals pay a cost for choosing an action different from the action of each of their friends.

and the distribution of preferences. The equilibrium concept is Pure Strategy Nash Equilibrium (PSNE). A profile of strategies $\mathbf{x}^*$ is a PSNE if it satisfies the standard requirement,[3]

$$\forall i, \forall x_i, \quad U_i\left(x_i^*, \mathbf{x}_{-i}^*, g\right) \geq U_i\left(x_i, \mathbf{x}_{-i}^*, g\right)$$

2.3 Benchmark: Continuous Social Norms

If the choice set is continuous on the positive real line, i.e., $x_i \in \mathbb{R}_+ \forall i$, Ushchev and Zenou (2020) (see Proposition 1 in the paper) prove there exists a unique interior PSNE $\mathbf{x}^*$, which is given by:

$$\mathbf{x}^* = (1-\lambda)(\mathbf{I} - \lambda\widehat{\mathbf{G}})^{-1}\alpha = \widehat{\mathbf{M}}\boldsymbol{\alpha} \tag{2.3.1}$$

where $\lambda \equiv \frac{\theta}{1+\theta}$ is a monotone transformation of the society's taste for conformity θ, $\boldsymbol{\alpha}$ is the vector of the society's (exogenous) preferences, and each element of the matrix of marginal effects $\widehat{\mathbf{M}} \equiv (1-\lambda)(\mathbf{I} - \lambda\widehat{\mathbf{G}})^{-1}$ is decomposed into a series (i) whose coefficients are given by a geometric distribution with odds ratio θ, and (ii) whose k^{th} term is proportional to the (normalized) number of paths from i to j of length k in the network g. In particular, $\widehat{m}_{ij}$ has the following form:

$$\widehat{m}_{ij} = \sum_{k=0}^{\infty}(1-\lambda)\lambda^k \widehat{g}_{ij}^{[k]}$$

Therefore, in equilibrium, each agent's optimal action x_i^* is a combination of her preference α_i and the preferences α_j of the other individuals in g, weighted by their proximity in the network. In addition, it is possible to express i's optimal action as a convex combination of her *preferences*, α_i, and her *endogenous social norm*, $\overline{x}_i^*$, which is given by equation (2.3.2) below. Precisely, Ushchev and Zenou (2020) show the i's action x_i^* is above (below) her social norm $\overline{x}_i^*$ if her preference parameter α_i is higher (smaller) than the weighted average of the other individuals' preferences (see Lemma 1 in the paper).

$$\overline{\mathbf{x}}^* = \widehat{\mathbf{G}}\widehat{\mathbf{M}}\boldsymbol{\alpha} = (1-\lambda)\sum_{k=0}^{\infty}\lambda^k \widehat{\mathbf{G}}^{k+1}\boldsymbol{\alpha} \tag{2.3.2}$$

In the rest of the paper, I relax the convenient assumption of a continuous choice set, analyzing which results reported in this section - and broadly, on Ushchev and Zenou (2020) - change or remain true.

[3] This PSNE definition is valid when the choice set is continuous or discrete.

2.4 Discrete Social Norms

The literature on conformity in network games mainly focuses on analyzing continuous actions. However, some decisions closely related to social norms are, by nature, discrete; for example, female labor force participation - Bursztyn et al. (2018) - or female genital cutting - Gulesci et al. (2021). In this section, I extend the benchmark to analyze a discrete choice model, shedding light on such scenarios. In particular, I assume individuals can choose between $k < \infty$ potential actions, that is, $x_i \in \{x^1, ..., x^k\} \quad \forall i$. For simplicity, I refer to this setting as the Discrete Social Norms Game (DSNG).

When the choice set is discrete, the equilibrium analysis of network games becomes challenging. Two forces are conditioning individual decisions: their preferences and local social norms. In the continuous case, the equilibrium outcome is a weighted average of both - see equation (2.3.1). However, in the discrete scenario, multiple equilibria generally arise, encompassing equilibrium outcomes in which either conformity or preferences prevail, or both are present. This feature is a direct consequence of DSNG being a supermodular game, as formally stated in the following proposition.

Proposition 1. The game $[(n, g), (x_1, ..., x_n; U_1, ..., U_n)]$ in which, $\forall i$, utility function is given by equation (2.2.1), and $x_i \in \{x^1, ..., x^k\}$, with $k < \infty$ is a supermodular game.
Proof: See Appendix 2.A.1.

Supermodular games have nice properties, helping the characterization of the set of PSNEs - Robinson (1951), Topkis (1979), Milgrom and Roberts (1990), Vives (1990). First, the mentioned set is non-empty: the existence of (at least one) PSNE is guaranteed. Second, the set of PSNEs has a greatest and a least element (GE and LE, respectively). The greatest (least) element of the set of PSNEs is the equilibrium such that individual actions are maximal (minimal), i.e., there is no other equilibrium in which an individual chooses a higher (lower) action. It follows that equilibrium is unique if and only if the greatest and least element coincide.

Importantly, the existence of complementarity in actions helps to develop an iterative procedure to find any equilibrium, as the direction of a potentially profitable deviation is pinned down by this complementarity. The algorithm to find the greatest and least equilibrium, consisting of iteratively applying the Best Response Function (BRF) of the game, is attributed to Robinson (1951) and Topkis (1979). Echenique (2007) presents an algorithm to find all the PSNEs in supermodular games, also iteratively applying BRF, but on sub-games in which individuals' actions are restricted. Throughout this section, I rely on these algorithms to characterize the set of equilibria of the DSNG.

Before proceeding, the next definition introduces the functional form of an individual profitable deviation in the DSNG, which provides useful intuition about the main results of this paper.

Definition 1. Let $[(n, g), (x_1, ..., x_n; U_1, ..., U_n)]$ be the game in which, $\forall i$, utility function is given by equation (2.2.1), and $x_i \in \{x^1, ..., x^k\}$, with $k < \infty$. For any pair of strategies $(x, x^{'})$ such that, WLOG, $x > x^{'}$, and any individual i, a Profitable Deviation (PD) for individual i from strategy $x^{'}$ to strategy x exists if the following condition holds,

$$\begin{aligned} PD_i(x, x^{'}) &\equiv U_i(x, \mathbf{x}_{-i}, g) - U_i(x^{'}, \mathbf{x}_{-i}, g) = \\ &= \left(\alpha_i - \frac{x + x^{'}}{2}\right) + \theta \left(\overline{x}_i - \frac{x + x^{'}}{2}\right) > 0 \end{aligned} \tag{2.4.1}$$

Formula's derivation: See Appendix 2.A.1.

Then, the existence of a profitable deviation for individual i depends on how the average of these actions $(x, x^{'})$ relates to her preferences, α_i, and her social norm, $\overline{x}_i$. If the two terms of $PD_i(x, x^{'})$ have the same sign, there is no trade-off between conformity and preferences for individual i when choosing between x and $x^{'}$. This is not true if i's preferences and social norms are oppositely related to the average action. In that case, the value of θ, the taste for conformity, weights social and private costs of choosing an action.

2.4.1 Two actions space

The simplest version of the DSNG is the binary action case. Precisely, suppose that $\forall i$, $x_i \in \{x^l, x^h\}$ where $x^l < x^h$. In this case, equation (2.4.1) pins down the optimal solution for any individual i, as she must compare only two actions. Indeed, i's optimal action is $x_i^* = x^h$ if $PD_i(x^h, x^l) > 0$ and $x_i^* = x^l$ if $PD_i(x^h, x^l) < 0$.

As previously stated, $PD_i(x^h, x^l)$ depends on i's preferences and social norm, and precise values of θ and (x^l, x^h). Thus, it is natural to argue that the equilibrium outcome, determined by a system of equations $PD(x^h, x^l)$, will depend on the distribution of preferences, the network structure (determining local social norms), and the parameters θ and (x^l, x^h). The interplay of these factors will pin down three types of equilibria featuring different degrees of *individualism* and *conformism.* The types of PSNE are defined below.

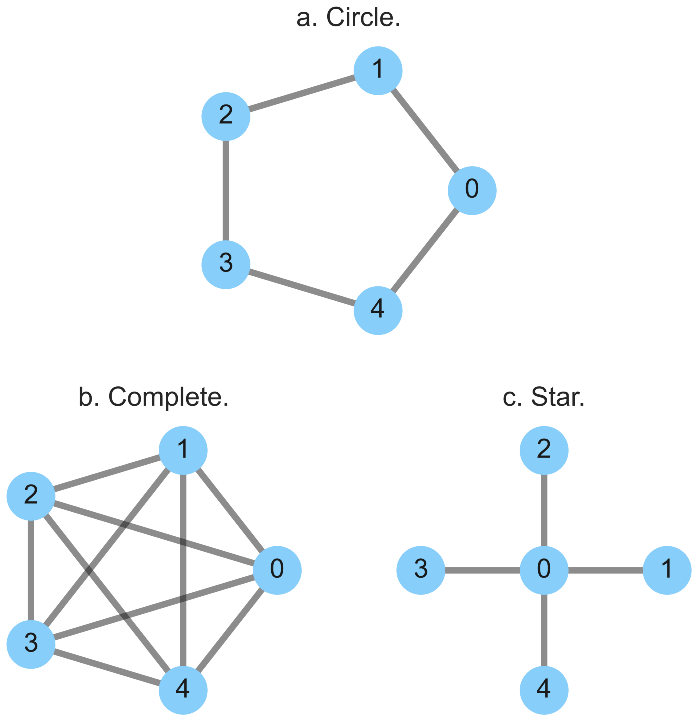
Definition 2. A *purely conformist equilibrium* (PC) is such that all individuals choose the same action, independently of their preferences. A *pure individualist equilibrium* (PI) is such that all individuals choose the closest action to their preferences, independently of their social norms. A *partially conformist-individualist* *equilibrium* (CI) is such that a subset of individuals conform to their social norms while others choose the closest action to their preferences. In the latter, the *degree of conformism (individualism)* at equilibrium refers to the size of the subset of conformist (individualist) agents.

The next proposition characterizes the set of PSNE for the DSNG when the choice set $\{x^l, x^h\}$ exhibits extremism, in the sense that $\forall i$, $x^l \leq \alpha_i \leq x^h$.

Proposition 2. Let $[(n, g), (x_1, ..., x_n; U_1, ..., U_n)]$ be the game in which, $\forall i$, utility function is given by equation (2.2.1) and $x_i \in \{x^l, x^h\}$. Let $\underline{\alpha} \equiv min_i \alpha_i$ and $\overline{\alpha} \equiv max_i \alpha_i$, and assume $\underline{\alpha} \geq x^l$ and $\overline{\alpha} \leq x^h$. If $\theta \geq 1$, then the game admits multiple equilibria. The greatest (least) element of the set of PSNEs exhibits *pure conformism*, in the sense that all individuals choose action x^h (x^l).
Proof: See Appendix 2.A.1.

Several comments are in order. First, when the society's taste for conformity is relatively high, $\theta \geq 1$, a purely conformist equilibrium exists, independently of the network structure and distribution of preferences. Other equilibria also exist, but nothing prevents society from falling into a purely conformist scenario. Second, if $\theta < 1$, generally, a unique equilibrium exhibiting high levels of *individualism* exists. In that case, the equilibrium corresponds to either a pure individualist or a partially conformist-individualist class. The degree of conformism depends, first, on the network structure and, second, on the distribution of preferences.

Figure 2.1 Example of a circle, complete, and star network with five nodes.



Consider the three network graphs in Figure 2.1 to gain intuition about this. These are three standard networks: a circle, a complete, and a star network. For the circle and complete networks and any distribution of preferences consistent with the assumption in Proposition 2, the unique equilibrium falls in the pure individualist class. However, the equilibrium outcome for the star network also depends on the specific distribution of preferences and the value of θ. For example, a partially conformist-individualist equilibrium arises if preferences are uniformly distributed on the interval $[x^l, x^h]$. Furthermore, the degree of individualism at equilibrium varies monotonically with the value of conformism: as θ approaches 1, the subset of individualist agents approaches 0.

2.5 Conclusion

This paper explores a model of social norms and conformity when individuals are embedded in a social network. Each individual selects a code of conduct based on personal preferences and the prevailing social norm. Social norms are local, endogenous, and discrete - aiming at understanding real-world phenomena like religious choice, female genital cutting, and stances on public policies such as abortion access or same-sex marriage.

As a first step of research, this paper shows that various social norms may emerge when actions are discrete due to the multiplicity of equilibria. Despite the local nature of the network interactions, global social norms arise in highly conformist societies, where all individuals choose the same code of conduct. If a unique equilibrium exists, it tends to exhibit a high degree of individualism, although certain individuals may still conform to the norm.

A natural extension of this project would study how precise choice set features - for instance, regarding its cardinality and whether actions are uniformly distributed or polarized - impact equilibrium outcomes, especially on the resulting social norms. In that regard, the analysis of this discrete action model in the limit and the comparison to its analogous continuous action model is left for future research.

In this paper, I identify four factors determining social norms: the society's distribution of preferences and taste for conformity, the network structure, and the available codes of conduct. A deep understanding of these and other forces behind endogenous social norms has theoretical and policy relevance.

2.A Appendix to Chapter 2

2.A.1 Proofs

Proof of Proposition 1. To prove that $[(n, g), (x_1, ..., x_n; U_1, ..., U_n)]$ is a supermodular game, it is sufficient to check it fulfills the following three conditions,

1. $\forall i$, the choice set $\{x^1, ..., x^k\}$ is a compact set.

2. $\forall i$, $U_i(x_i, \mathbf{x}_{-i}, g)$ given by equation (2.2.1) is a continuous function on $(x_i, \mathbf{x}_{-i})$.

3. $\forall(i, j), j \neq i$, $U_i(x_i, \mathbf{x}_{-i}, g)$ has increasing differences in (x_i, x_j).

It is straightforward to note that the first two conditions hold. To finish the proof, it remains to verify that condition 3. is fulfilled. To prove that $U_i(x_i, \mathbf{x}_{-i}, g)$ has increasing differences in (x_i, x_j), it is enough to show that, $\forall x \geq x^{'}$ and $\forall x_j$, $U_i(x, \mathbf{x}_{-i}, g) - U_i(x^{'}, \mathbf{x}_{-i}, g)$ is non-decreasing in x_j (see Levin (2003)), where,

$$U_i(x, \mathbf{x}_{-i}, g) - U_i(x^{'}, \mathbf{x}_{-i}, g) = (x - x^{'}) \left[\alpha_i - \frac{x + x^{'}}{2}\right] - \frac{\theta}{2}\left[(x - \overline{x}_i)^2 - (x^{'} - \overline{x}_i)^2\right]$$

If $g_{ij} = 0$, the above equation is independent of x_j, and the condition trivially holds. If $g_{ij} = 1$, then x_j is one of the terms appearing on i's social norm, $\overline{x}_i$. Thus, proving that $U_i(x, \mathbf{x}_{-i}, g) - U_i(x^{'}, \mathbf{x}_{-i}, g)$ is is non-decreasing in $\overline{x}_i$ is equivalent to proving the original condition. Furthermore, since $\overline{x}_i$ only appears in the second term of the equation, the condition holds if the auxiliary function $f(\overline{x}_i)$ defined below is non-increasing in $\overline{x}_i$,

$$f(\overline{x}_i) \equiv \left[(x - \overline{x}_i)^2 - (x^{'} - \overline{x}_i)^2\right] = x^2 - (x^{'})^2 - 2\overline{x}_i(x - x^{'})$$

Provided $x \geq x^{'}$, $f(\overline{x}_i)$ is non-increasing in $\overline{x}_i$ and, thus, the proof is complete.

Formula for Definition 1. The formula $PD_i(x, x^{'})$ is obtained by simple math:

$$\begin{aligned} PD_i(x, x^{'}) &\equiv U_i(x, \mathbf{x}_{-i}, g) - U_i(x^{'}, \mathbf{x}_{-i}, g) \\ &= \alpha_i(x - x^{'}) - \frac{x^2 - (x^{'})^2}{2} - \frac{\theta}{2}\left[(x - \overline{x}_i)^2 - (x^{'} - \overline{x}_i)^2\right] \\ &= (x - x^{'})\left[\alpha_i - \frac{x + x^{'}}{2}\right] - \frac{\theta}{2}\left[x^2 - (x^{'})^2 - 2\overline{x}_i(x - x^{'})\right] \\ &= (x - x^{'})\left[\left(\alpha_i - \frac{x + x^{'}}{2}\right) + \theta\left(\overline{x}_i - \frac{x + x^{'}}{2}\right)\right] \end{aligned}$$

Provided $x > x^{'}$, a Profitable Deviation from strategy $x^{'}$ to strategy x exist if:

$$PD_i(x, x^{'}) = \left(\alpha_i - \frac{x + x^{'}}{2}\right) + \theta\left(\overline{x}_i - \frac{x + x^{'}}{2}\right) > 0$$

Proof of Proposition 2. Let $[(n, g), (x_1, ..., x_n; U_1, ..., U_n)]$ be the game in which, $\forall i$, utility function is given by equation (2.2.1) and $x_i \in \{x^l, x^h\}$. Let $\overline{\alpha} \equiv max_i \alpha_i$ and $\underline{\alpha} \equiv min_i \alpha_i$, and assume $\overline{\alpha} \geq x^h$ and $\underline{\alpha} \leq x^l$.

Let me first show that, provided $\theta \geq 1$, the greatest equilibrium (GE) is such that, $\forall i$, $x_i = x^h$. The action x^h is optimal for individual i if and only if $PD_i(x^h, x^l) \geq 0$, where,

$$\begin{aligned} PD_i(x^h, x^l) = \left(\alpha_i - \frac{x^h + x^l}{2}\right) + \theta\left(\frac{x^h - x^l}{2}\right) \geq 0 \\ \Leftrightarrow \quad \alpha_i \geq \frac{x^h + x^l}{2} - \theta\left(\frac{x^h - x^l}{2}\right) \end{aligned}$$

the first line uses the fact that, under GE, $\overline{x}_i = x^h$ $\forall i$. The individual i who is most likely to deviate from choosing action x^h has the smallest preference parameter, $\underline{\alpha} \equiv min_i \alpha_i$. No one deviates from GE if such an individual does not do it. Then, the strategy profile such that $\forall i$, $x_i = x^h$ is the GE if

$$\begin{aligned} & \underline{\alpha} \geq \frac{x^h + x^l}{2} - \theta\left(\frac{x^h - x^l}{2}\right) \\ \Leftarrow \quad & x^l \geq \frac{x^h + x^l}{2} - \theta\left(\frac{x^h - x^l}{2}\right) \\ \Leftrightarrow \quad & 0 \geq (1 - \theta)\left(\frac{x^h - x^l}{2}\right) \\ \Leftrightarrow \quad & \theta \geq 1 \end{aligned}$$

where the second line relies on the assumption that $x_l \leq \underline{\alpha}$, and the last line uses the fact that $x^h - x^l > 0$. Analogously, one can prove that, provided $x_h \geq \overline{\alpha}$, the least equilibrium (LE) is such that $\forall i$, $x_i = x^h$. The multiplicity of equilibria follows directly from the observation that the greatest and least elements of the set of PSNEs are not equal.

Chapter 3

Hate Speech and Social Media: Evidence from Bolsonaro's Election in Brazil

Coauthored with Diego Marino Fages.

3.1 Introduction

Which factors influence individuals' perception of social norms? Do these perceptions translate into behavior? What happens when these factors undergo a sudden change? This paper sheds light on these questions by exploring how the arrival of new *political information* triggers a change in *social norms* and, consequently, affects individuals' *expressions of opinions*. Specifically, we provide novel evidence on how the 2018 presidential election of Jair Bolsonaro in Brazil affected individuals' online expressions of hate.

Two key assumptions underlie our research question. Firstly, we consider the 2018 election result to be an information shock, that is, new and potentially unexpected political information. The evidence suggests this is a realistic assumption, as Bolsonaro's election surprised the Brazilian community. Bolsonaro got 46% of the votes in the 1º round of the election and 55% in the 2º round. The opinion polls conducted by diverse companies in the days before the election estimated that Bolsonaro's vote share would be approximately 35% for the 1º round, and only one polling company estimated a vote share above 40% of the votes.[1]

Second, in line with the claim by Bursztyn et al. (2020), we posit that the election result may trigger a quick update of the prevailing social norm governing what types of speech are socially acceptable. Bolsonaro, sometimes called "the Trump

[1] Source: Wikipedia, access date: June 2023.

of the Tropics," is widely recognized for his contentious viewpoints, encompassing homophobia, racism, and sexism.[2] Consequently, armed with the knowledge that most of the population voted for Bolsonaro, individuals may reassess their perception of the social acceptability of such controversial rhetoric. Under the premise that this perception was not entirely accurate prior to the elections, we might expect a behavior change after it.

In this project, we rely on data from the social media platform Twitter. To conduct the empirical analysis, we build a longitudinal dataset of tweets spanning the period between July 2017 and December 2019. This time frame covers approximately one year leading up to the electoral rally and another year following the assumption of office by the 38th Brazilian president. We combine the data we retrieve from Twitter with the 2018 election results at the municipality level, whose data source is the Superior Tribunal Court (TSE), and with geospatial data from the Brazilian Institute of Geography and Statistics (IBGE) to geo-locate tweets and election outcomes. Finally, we create two datasets, derived from the original tweets' dataset, for the empirical analysis. In the first dataset, the cross-sectional unit is a Brazilian municipality, and the time unit is a day. The corresponding units in the second longitudinal dataset are a Twitter user and a month.

Our primary variable of interest is the daily (monthly) frequency of hate speech within each Brazilian municipality (Twitter user). To construct it, we downloaded a representative sample of the universe of tweets in Portuguese that provide geolocation in Brazil. After cleaning the data, we process the text of each tweet to determine if they contain hate speech. To accomplish this, we rely on text analysis techniques.[3] Specifically, we fine-tune[4] a pre-trained *Bidirectional Encoder Representations from Transformers* (BERT) model to be suitable for the hate speech detection task. Our classification model was trained using the Portuguese BERT model introduced by Souza et al. (2020) and the hate speech dataset presented by Fortuna et al. (2019).

To identify the impact of the 2018 election of Bolsonaro on hate speech, we propose two difference-in-differences design models. In the first model, a traditional difference-in-differences, we split municipalities into control and treatment groups according to the vote share received by Bolsonaro in the 1º round of the election. Specifically, any municipality in which Bolsonaro's vote share is lower than the national outcome, 46% of the votes, falls into the treatment group. For the second

[2] To illustrate this point, consider a sample of Bolsonaro's statements: "*I would be incapable of loving a homosexual son*," "*The scum of the earth is showing up in Brazil as if we did not have enough problems of our own to sort out*," and (speaking to a Brazil Congresswoman) "*I would not rape you because you do not deserve it*." Sources: CNBC web portal, Reuters, AP News, and USA Today. Access date: June 2023.

[3] See Gentzkow et al. (2019) and Ash and Hansen (2023) for surveys on text-as-data and economics.

[4] Fine-tuning is the technique of training a pre-trained model on a suitable dataset for a new task. In our case, this new task is hate speech detection.

model, similarly to Albornoz et al. (2022), we propose a difference-in-differences design with a continuous treatment variable - see Callaway et al. (2021) for a theoretical reference. In this case, the treatment variable is Bolsonaro's vote share in each Brazilian municipality, which measures the local incidence of the information shock, i.e., the 1^{o} round election outcome.

We find that online hate speech increased after the 2018 presidential election. At the municipality level, this increase is mainly driven by regions where Bolsonaro *lost*. Furthermore, our findings suggest that the magnitude of the information shock, i.e., the election results, is crucial to explaining the extent of the rise in hate expressions. The largest increase in hate speech is observed in municipalities where Bolsonaro was particularly unpopular. As Twitter data allows us to analyze individual data, we further explore *who* is driving this result. Our results at the individual level indicate that both the intensive and extensive margins of hate speech contributed to this explain this phenomenon. Some Twitter users started to post hate speech tweets after the elections, especially in the municipalities where Bolsonaro lost. In addition, users who posted tweets with hate content increased this behavior's frequency after the elections.

We interpret these findings through the lens of a belief update mechanism. Following the information shock triggered by the 2018 election result, individuals living in a relatively against-Bolsonaro municipality could revise their beliefs regarding socially acceptable speeches. Once the social norm is updated, these individuals may feel justified in expressing controversial and hateful viewpoints through social media platforms, even if they reside in a municipality where the pre-election prevalence of such behavior was relatively low.

Related Literature. We contribute to the economic literature that studies the impact of political information on social norms, particularly the literature documenting that political changes lead to fast changes in social norms and behavior. Bursztyn et al. (2020) run two experiments linking the rise of Donald Trump's popularity in the US and the social acceptability of xenophobia. In the first experiment, the authors document that Trump's victory increased individuals' willingness to express xenophobic opinions. The second experiment focuses on sanctioning xenophobic expressions; the results show that these expressions are less likely to be sanctioned in an environment where those views are relatively more popular. Albornoz et al. (2022) argue that the Brexit referendum caused a shift in the social acceptability of xenophobic expressions. The authors show that hate crime increased after the referendum, especially in areas with a larger share of "remain" votes. The authors interpret these results within a framework of conformity and misperception of the prevailing social norm at the national level so that when the referendum results arrive, a social norm update impacts the expression of anti-immigrant attitudes.

In broader terms, our paper speaks to the economic literature on social norms and conformity. We analyze the effects of a social norm update, departing from the literature that studies social norms persistence[5] - among others, Giuliano (2007), Fernandez (2007), and Alesina et al. (2013). In addition, our paper contributes to the literature that examines the interplay between norms and political institutions (Acemoglu and Jackson (2017)) or behaviors (e.g., Gerber et al. (2008), DellaVigna et al. (2016), and Perez-Truglia and Cruces (2017)). Finally, our paper connects with the literature on social norms by studying their geographical variation within a country and, furthermore, by analyzing high-frequency individual-level data.

This paper speaks to the literature linking social media and expressions of hate, particularly against minority groups.[6] Müller and Schwarz (2023) find a positive relationship between Twitter usage and ethnic hate crimes since the presidential election of Donald Trump in the US, pointing out that social media may enable people with extreme viewpoints to find a source of legitimacy. Bursztyn et al. (2019) show that social media increased ethnic hate crimes in Russian cities with high pre-existing anti-immigrant sentiments. Müller and Schwarz (2021) find evidence that social media affects the propagation of anti-refugee incidents in Germany. Focusing on sex crime, Bhuller et al. (2013) document an increase in this type of crime associated with the roll-out of broadband internet in Norway. This piece of literature covers several social media platforms, like Twitter and Facebook, but focuses mainly on xenophobia and ethnic hate crimes. In this paper, we consider a wider definition encompassing different hate targets. In contrast to the existing literature, this paper focuses on hate speech rather than hate crime and online rather offline expressions of hate.

The rest of the paper is organized as follows. Section 3.2 describes the data. Section 3.3 presents the identification strategy, and section 3.4, the results at the municipality and individual levels. Section 3.5 concludes.

3.2 Data

In this paper, we aim to understand how the 2018 presidential election of Bolsonaro affected online hate speech in Brazil. Our primary data source is the social media platform Twitter, from which we measure online hate speech at the municipality level and in the period under study. We combine the data we retrieve from Twitter with three types of administrative data. First, we use the 2018 election results at the municipality level, whose data source is the Superior Tribunal Court (in Portuguese, *Tribunal Superior Eleitoral* - TSE), the highest structure within the Brazilian Electoral Justice system. In addition, we rely on geospatial data from the

[5] See Bisin and Verdier (2011) for a survey.

[6] In addition to this literature, other research has linked diverse types of traditional media to violence - Dahl and DellaVigna (2009), Card and Dahl (2011), Yanagizawa-Drott (2014), DellaVigna et al. (2014), and Ivandic et al. (2019), among others.

Brazilian Institute of Geography and Statistics (in Portuguese, *Instituto Brasileiro de Geografia e Estatística* - IBGE) to geo-locate tweets and election results. Lastly, we use the 2010 Population Census in Brazil microdata from IBGE to construct demographic variables aggregated at the municipality level.

3.2.1 Twitter data

Twitter is an online platform that allows users to publish short messages, of a maximum of 140 characters, on their profiles. With one of the largest Twitter user bases in the world, Brazil is an appealing case of study for online activity - in this case, related to Twitter users' speech. In January 2022,[7] Brazil ranked fourth worldwide regarding the number of Twitter users, with an estimated 19 million active accounts (after the United States, Japan, and India). Importantly, most of the Brazilians who were online in 2022 used social media for news (64%)[8] and political discussion (78%),[9] which are closely related to this paper's topic. Another advantage of this study case is that online hate speech, as opposed to hate crime, can be directly observed and quantified - in this case, by analyzing tweets' content. Thus, online hate speech is not subject to changes in reporting.[10]

In the empirical analysis, our main variable of interest is the proportion of tweets classified as hate speech per municipality (or individual) and date. The next paragraphs describe how we collected and processed Twitter data to construct this variable.

Data collection. We use the Twitter Application Programming Interface v2 (Twitter API v2) to collect our data. Specifically, we rely on the *v2 full-archive search endpoint*, which gives access to the entire history of publicly available (and yet undeleted) tweets. We retrieve all the tweets, net of retweets, which satisfy three conditions specified in the Twitter query. First, tweets must be written in Portuguese. Second, tweets must provide geo-location information and be located in Brazil. Lastly, tweets must belong to the period comprised between July 2017 to December 2019, both included. As the daily amount of data retrieved by this query is around 300.000 tweets, we further restrict the Twitter query to retrieve only tweets posted on any Monday belonging to the mentioned period. This query imposes two main assumptions on our tweets' sample. We assume that (i) tweets posted on any Monday and (ii) geo-located tweets are representative samples of the tweets'

[7] Source: Statista web portal, access date: June 2023.

[8] Source: Digital News Report, 2022, Reuters Institute & University of Oxford, access date: June 2023.

[9] Source: Statista web portal, access date: June 2023.

[10] Online hate speech also differs from hate crime regarding its cost and timing. In the former, the perpetrator immediately pays the cost of expressing hateful content. On the other hand, a hate crime must be reported and processed by justice before the perpetrator pays its cost.

universe. Appendix 3.A.1 provides supportive evidence for these assumptions and complementary information to this section.

Data processing. We extract relevant content from the tweets' text, which will serve as input for the hate speech detection task. We exclude punctuation marks, stop-words, and multimedia items. We do not remove negative stop-words that may change the statement's meaning: *"mas"* (but), *"nem"* (neither), *"não"* (no), *"sem"* (without), and *"fora"* (out). We anonymize user mentions and URL links but keep hashtags in their native Twitter format, as they may contain relevant information. We drop all tweets containing only links and/or user mentions and those posted by accounts created after 2018. The reason for the latter is to exclude from the analysis accounts potentially created in the context of the electoral rally.

Hate speech detection. We rely on Natural Language Processing (NLP) techniques to detect hate speech in our tweets' sample. We train a pre-trained *Bidirectional Encoder Representations from Transformers* model (BERT model) on a dataset specific to the hate speech detection task. This process is known as *fine-tuning* a pre-trained model. Specifically, we use *BERTimbau*, a BERT model for Brazilian Portuguese by Souza et al. (2020), and train it on a dataset of tweets in Portuguese, by Fortuna et al. (2019).

Souza et al. (2020) present *BERTimbau*, a BERT model for Brazilian Portuguese, in two sizes, Base and Large. In this paper, we fine-tune BERTimbau-Base for the hate speech detection task. Its architecture comprises 12 layers, 768 hidden dimensions, 12 attention heads, and 110M parameters. The authors trained their model on the *brWaC corpus* by Wagner Filho et al. (2018) and two NLP tasks, Masked Language Modeling (MLM) and Sentence Prediction (NSP).

In their paper, Fortuna et al. (2019) collected 5668 tweets in Portuguese through Twitter API from January to March 2017. The authors provide two annotation schemes for the dataset, a binary, and a hierarchical multiple classifications. Each tweet classified as "hate speech" is further split into classes for the hierarchical classification. Its second-level classification relates to the target of hate, and it comprises: "sexism," "body," "origin," "homophobia," "ideology," "religion," "health," and "lifestyle." As a first step, this paper uses the binary classification dataset to fine-tune the mentioned BERT model, in which 31.5% of the tweets were annotated as "hate speech." To construct it, three (Portuguese native) annotators labeled every tweet as "hate speech" or "not hate speech," and the authors applied the majority vote to determine the final annotation of each tweet.

Before fine-tuning, we divide the dataset between 80% for training, 10% for validation, and 10% for testing. In NLP applications, the performance of a model in a given task is directly influenced by the characteristics of the training sample. In the case of Fortuna et al. (2019)'s dataset, as in other datasets on hate speech detection, a class imbalance exists. Tweets annotated as "hate speech" constitutes the minority class. As class imbalance may affect a model's performance in a text classification task, we use a Random Oversampling technique[11] to equalize the number of tweets per class in the training sample. Our model attains an overall accuracy of 77% in both the validation and test samples. Appendix 3.A.2 provides further details on the hate speech detection task, resources utilized, and the model's training results.

Data classification. After training the BERT model for the hate speech detection task, we use it to detect hate speech in the tweets on the sample. We construct a binary variable 0/1, named "predicted hate speech," as a result of this classification. Then, we use the tweet-specific geo-location information to map each tweet to the Brazilian municipalities based on latitude and longitude through IBGE's geospatial shape files. Finally, we compute the proportion of tweets containing hate speech by municipalities (or individuals) over time, which is the main outcome variable of this paper.

3.2.2 Administrative data

The election result we use as an information shock is the vote share obtained by Bolsonaro in the 1^{o} round of the 2018 Brazilian presidential election. The Superior Tribunal Court (TSE) provides official data at the municipality level on all election results in Brazil since 1994. Given that TSE's records do not contain the geo-coordinates of the electoral districts, we rely on geospatial data from the Brazilian Institute of Geography and Statistics (IBGE) to determine their location. IBGE provides Brazilian geospatial data at country, state, and municipality levels. Lastly, we use microdata from the 2010 Population Census in Brazil, the last available for the pre-Bolsonaro period. Consistently with our analysis unit, we aggregate the census microdata at the municipality level.

3.2.3 Descriptive statistics

We study how the 2018 Brazilian presidential election influenced online hate speech. To accomplish this, we create two longitudinal datasets of geo-located tweets spanning from July 2017 to December 2019.

In the first dataset, the time unit is a day t - for any Monday included in the tweets' sample - and the cross-sectional unit is a Brazilian municipality m. The main variable is the proportion of tweets classified as hate speech for a given date t

[11] Random oversampling involves transforming the existing data to adjust the class distribution. It consists of randomly selecting examples from the minority class and adding them to the original dataset.

and municipality m. Brazil is divided into twenty-six states and one federal district. Each sub-national entity is further divided into municipalities, and Brazil currently has 5570 municipalities. In the empirical analysis, we use data from approximately 1500 municipalities for which Twitter data is available after data cleaning.[12] The longitudinal dataset at the municipality level is unbalanced, with some municipalities present over the period and others for which Twitter data is relatively more scarce. On average, we observe each municipality on approximately 100 Mondays (with a standard deviation of 37 days).

In the second longitudinal dataset, the time unit is a month t, and the cross-sectional unit is a Twitter user i. We include any Twitter user whose tweets are geo-located in no more than two different municipalities. The longitudinal dataset at the individual level is also unbalanced, as Twitter activity significantly varies for different individuals. In the regression analysis, we further restrict our attention to the sub-sample of users (i) who posted tweets in the pre and post-election periods and (ii) such that we observe at least 5 tweets per user per month. On average, we observe 190 tweets for each Twitter user distributed over approximately 12 months (6 months before and 6 months after elections).

This paper builds upon two fundamental observations. Firstly, we acknowledge that the presidential election, which we consider an information shock, did not uniformly affect all Brazilian citizens. Instead, we observe a geographical variation in Bolsonaro's vote share, which helps us to identify the effect of interest. Secondly, the evolution of online hate speech was not consistently constant throughout the period.

Regarding the first observation, Figure 3.1 shows that Bolsonaro's popularity varied across states and municipalities. Specifically, Bolsonaro's vote share was between 3% and 79% in the 1º round of the 2018 presidential election, which is the result we use in our empirical strategy to measure the information shock. As can be seen, the corresponding map for the 2º round results shows a similar geographical pattern. In Appendix 3.A.3, we present the (bimodal) distributions of these vote shares at the municipality level.

As for the second observation, Figure 3.2 shows the proportion of Brazilian tweets classified as hate speech in the period under study. The solid line corresponds to the raw data, consisting of the daily proportion of hate speech tweets, whereas the dotted line corresponds to the monthly average of hate speech. The shadow areas in the graph delimit the periods in which (i) the Presidential Election took place and (ii) Bolsonaro took office.[13] As can be seen, there was a sharp increase in hate speech during this period. The hate speech peaks on the data correspond to the closest

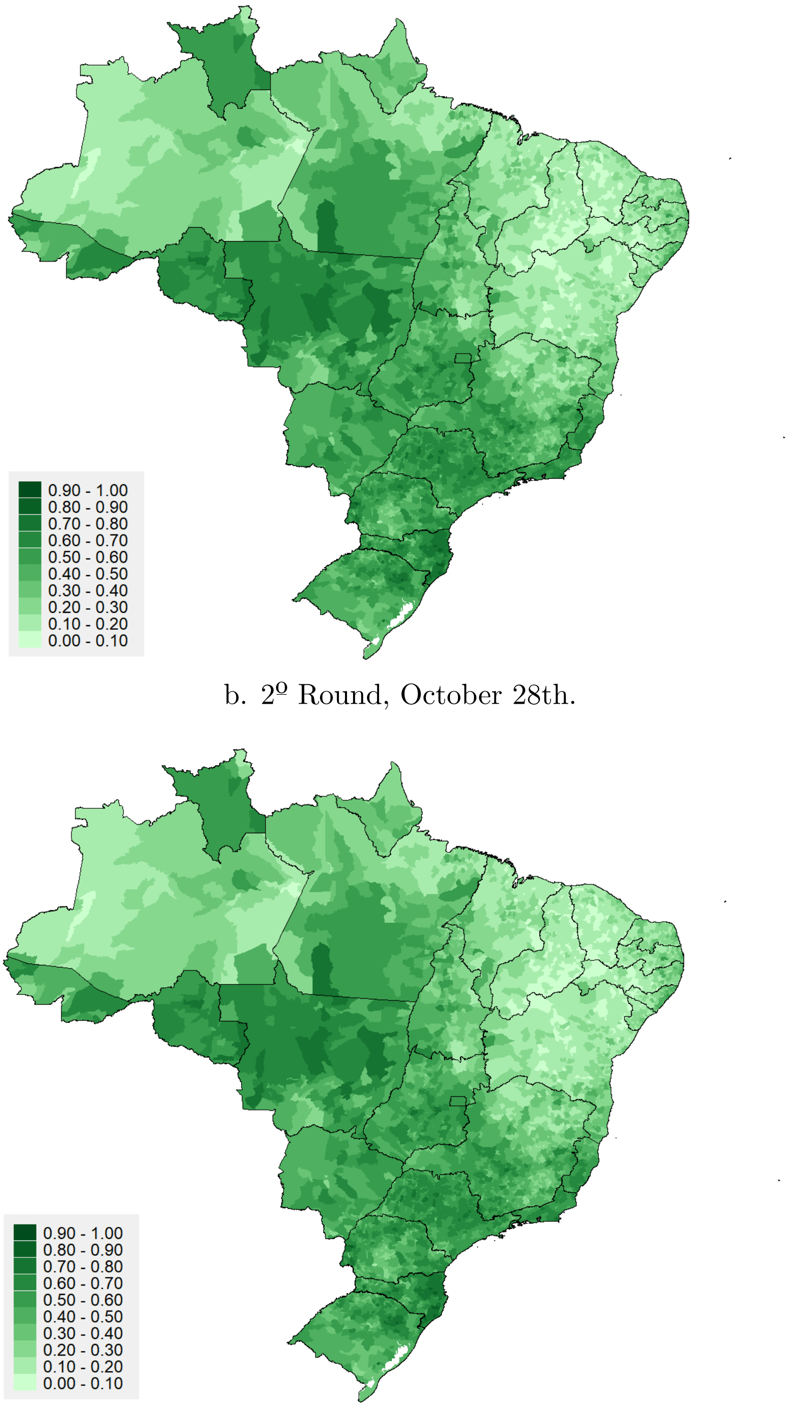
[12] We include in the empirical analysis any municipality for which we observe (i) at least 10 tweets daily and (ii) at least 10 times during 2017-2019.

[13] Specifically, the first and 2º rounds of the presidential election took place on October 7th and 28th, respectively. Bolsonaro took office as Brazil's 38th president on January 1st, 2019.

(but later on time) date in our sample to the first and 2º rounds of the election.[14]

Figure 3.1 Bolsonaro's vote share at the municipality level.

a. 1º Round, October 7th.

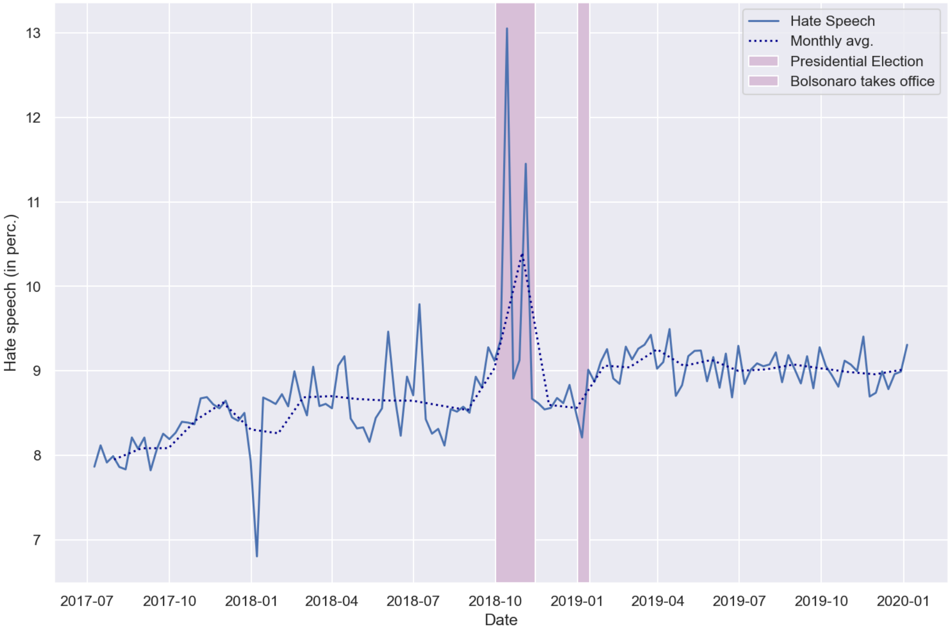


b. 2º Round, October 28th.

[14] There exist two other (although smaller) peaks in the data, during June and July 2018, corresponding to dates when Brazil's football team played a match in the 2018 World Cup. Figure 3.7 in Appendix shows that these peaks also correspond to a sharp increase in Twitter activity. Specifically, the daily amount of tweets is around a 50% higher during the period relative to the average. It is also worth noticing that the period with lower levels of hate speech corresponds to dates around the 2018 New Year break. Remarkably, this sharp decrease in hate speech was not observed around the 2019 New Year break, as the date coincides with when Bolsonaro took office.

Importantly, Figure 3.2 reveals that hate speech through Twitter increased post-election, i.e., after October 2018. The average proportion of hate speech from July 2017 to July 2018 was 8%, whereas it was 9% from January to December 2019.[15] Note that the above figure is constructed by aggregating hate speech at the national level, so it does not explore the sub-national evolution of hate speech over the period. The rest of this paper aims to answer whether this evolution was uniform (or not) across municipalities and why.

Figure 3.2 Evolution of hate speech in Brazilian tweets, 2017-2019.



Note: The variable Hate speech (in percent) is, for each date, the ratio of tweets classified as hate speech over the total amount of tweets.

3.3 Empirical strategy

We aim to estimate the effect of Bolsonaro's election - and the electoral rally - on hate speech. In the previous section (see Figure 3.2), we showed that hate speech on Twitter increased after Bolsonaro's election in comparison to the pre-election period at the national level. However, this is not sufficient to conclude that his election is to blame. It is possible that the election result responded to the rise in hate speech or that some other social phenomena are causing both the increase in hate speech and the political movement to the right.

The fact that these are national elections leaves us with no clear control group where Bolsonaro is not elected for president. However, his popularity varies across

[15] Figure 3.11 in Appendix 3.A.3 supports this observation. The mentioned figure is analogous to the one presented in the main text but with a standardized variable. As can be seen, almost all data points are below zero in the pre-election period and above zero in the post-election period.

states and municipalities (see Figure 3.1). We can then follow Albornoz et al. (2022) and exploit this differential informational shock to study whether hate speech increased relatively more in some places than others. First, we separate the municipalities based on the results of the 1º round of the elections: those where Bolsonaro got *at least* or *at most* the percentage of votes he got at the national level, 46%. For the sake of simplicity, we say that Bolsonaro "lost" the 1º round of elections (or simply, lost) in a municipality if his vote share was lower than 46%. Otherwise, we say that Bolsonaro "won" the election in that municipality. Thus, we perform a difference-in-differences analysis. Formally, we regress,

$$Hate_{mt} = \alpha_0 + \alpha_1 * Post_t * Lost_m + \delta_t + \pi_m + \epsilon_{mt} \tag{3.3.1}$$

where $Hate_{mt}$ is the share of tweets that contain hate speech in municipality m and date t, $Post_t$ is a dummy variable that takes the value one after the elections, $Lost_m$ is a dummy variable that takes the value one for the municipalities where Bolsonaro lost the elections (that is, his vote share was lower than 46%), δ_t is a linear time trend, π_m are municipality fixed effects, and ϵ_{mt} is a municipality-time specific error term. In this case, the identifying assumption is the traditional parallel trends assumption. That is, in the absence of the information shock, the difference in hate speech between municipalities where Bolsonaro won and lost the elections is constant over time.

Since our rich dataset allows us to follow Twitter accounts over time, we can further analyze hate speech at the individual level. Indeed, the availability of data at the individual level is an advantage of this paper, compared to Albornoz et al. (2022) and Carr et al. (2020), who studied hate crime at a more aggregate level. The purpose of the individual-level regressions is twofold. Firstly, it allows us to rule out the possibility that the rise in hate speech is driven by a change in the composition of the users before and after the elections. Secondly, individual data allows us to explore the intensive and extensive margins of hate speech. In other words, is the increase in hate speech driven by people already tweeting hate content before the elections, i.e., intensive margin, or is it caused by people who had not tweeted hate content before, i.e., extensive margin? Formally, we regress,

$$Hate_{imt} = \alpha_0 + \alpha_1 * Post_t * Lost_{im} + \delta_t + \gamma_i + \epsilon_{imt} \tag{3.3.2}$$

where $Hate_{imt}$ is the share of tweets that contain hate speech of account i in municipality m at month t, $Post_t$ is a dummy variable that takes the value one after the elections, $Lost_{im}$ is a dummy variable that takes the value one for the accounts located in municipalities where Bolsonaro lost the elections, δ_t is a linear time trend, γ_i are user fixed effects, and ϵ_{imt} is an account-municipality-time specific error term. In both cases, our coefficient of interest is α_1, which, given parallel trends, captures the *average treatment effect* (ATE).

Finally, we also exploit the continuous variation in Bolsonaro's vote share across municipalities. To do this, we replace $Lost_m$ in equation (3.3.1) and $Lost_{im}$ in equation (3.3.2) with the actual vote share Bolsonaro received in each municipality, $VoteShare_m$ and $VoteShare_{im}$. Formally,

$$Hate_{mt} = \beta_0 + \beta_1 * Post_t * VoteShare_m + \delta_t + \pi_m + \epsilon_{mt} \tag{3.3.3}$$

and,

$$Hate_{imt} = \beta_0 + \beta_1 * Post_t * VoteShare_{im} + \delta_t + \gamma_i + \epsilon_{imt} \tag{3.3.4}$$

where $Hate_{mt}$ ($Hate_{imt}$) is the share of tweets that contain hate speech in municipality m and date t (for user i in month t), $Post_t$ is a dummy variable that takes the value one after the elections, $VoteShare_m$ ($VoteShare_{im}$) is the share of votes obtained by Bolsonaro in municipality m (where individual i is located), δ_t is a linear time trend, π_m and γ_i are municipality and individual fixed effects, respectively, ϵ_{mt} is a municipality-time specific error term, and ϵ_{imt} is an account-municipality-time specific error term.

In both cases, our coefficient of interest is β_1. It captures the *average causal response* (ACR) on the treated to an incremental change in the dose, where the dose is the share of votes obtained by Bolsonaro in the municipality. The main identification assumption, in this case, is the strong parallel trends. It requires that, for all doses, the average change in hate speech over time across all municipalities that received a given dose is the same as the average change in hate speech that would have occurred over time for all municipalities that experienced a different dose - see Callaway et al. (2021).[16] Notice that, by definition, α_1 in equations (3.3.1) and (3.3.2) and β_1 in equations (3.3.3) and (3.3.4) have opposite signs: while the former captures the effect of $Lost_m = 1$, which depends negatively on Bolsonaro's vote share, the latter is proportional to it.

3.4 Results

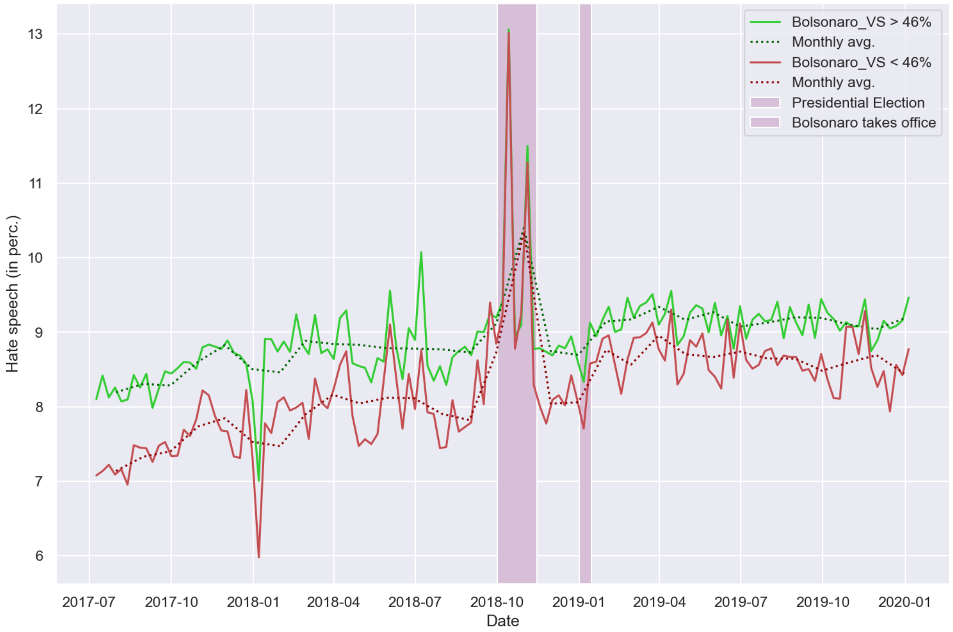
In this section, we present the main results of this paper. First, we document that hate speech increased after the 2018 presidential elections, especially in the municipalities where Bolsonaro lost. Then, we offer the results at the individual level, indicating that both the intensive and extensive margins of hate speech contributed to this explain this phenomenon.

[16] Formally, let d be the dose and Y_t be the potential outcome in time t. Then, the strong parallel trends assumption implies that for all d in D: $E[Y_t(d) - Y_{t\ 1}(0)] = E[Y_t(d) - Y_{t\ 1}(0)|D = d]$.

3.4.1 Municipality level

Before presenting the regression results, let us describe the municipalities that are in the treatment and control groups according to equations (3.3.1) and (3.3.2). Figure 3.3 below is an analogous figure to Figure 3.2, but now splitting the hate speech trends between treatment and control groups.[17]

Figure 3.3 Evolution of hate speech in Brazilian tweets, 2017-2019. Municipalities, by the 2018 election result.



The green lines correspond to the daily and monthly proportion of Brazilian tweets classified as hate speech for the municipalities in which Bolsonaro got at least 46% of the votes in the $1^{\underline{o}}$ round of the 2018 presidential election, i.e., where $Lost_m = 0$. On the contrary, the red lines correspond to the municipalities where Bolsonaro's vote share was at most 46%, that is, where $Lost_m = 1$. Again, the shadow areas in the graph delimit the periods in which the presidential election took place, and Bolsonaro took office.

Importantly for our identification strategy, the gap between hate speech pre-trends for treatment and control groups seems constant over time, i.e., pre-trends are parallel.[18] Furthermore, ratios of hate speech in municipalities where Bolsonaro won and lost seem to respond similarly to shocks; for example, both decreased around the 2018 New Year's Eve and increased during the 2018 World Cup (in July). Nonetheless, the hate speech trends in municipalities where Bolsonaro won and lost

[17] In Appendix 3.A.3, we present analogous graphs to Figures 3.2 and 3.3 but with hate speech aggregated at the monthly level. That is, only plotting the dotted lines in Figures 3.2 and 3.3.

[18] Figure 3.11 in Appendix 3.A.3 supports this observation. The mentioned figure is analogous to the one presented in the main text but with standardized variables. Once differences in levels are canceled, it is easy to see that the two lines move together.

changed after the election. As can be seen, the gap between hate speech ratios gets smaller, especially after Bolsonaro took office as the 38th Brazilian president. This evidence suggests a different reaction to Bolsonaro's election in one and another region.

Table 3.1 Descriptive Statistics, by the 2018 election result.

Variable	Won_m	Lost_m	Difference
urban	0.828	0.738	0.089***
income_pc	755.3	478.3	277.0***
cellphone	0.891	0.810	0.081***
computer	0.427	0.241	0.186***
internet	0.735	0.693	0.042***
primary	0.382	0.368	0.013***
tertiary	0.435	0.474	-0.039***
no_religion	0.058	0.070	-0.012***
catholic	0.678	0.721	-0.043***
pentecostal	0.128	0.117	0.010***
black	0.052	0.082	-0.029***
indigenous	0.003	0.006	-0.003**
brown	0.286	0.489	-0.202***
born_mun	0.567	0.673	-0.106***
born_state	0.673	0.747	-0.075***
vs_pt_2006	0.351	0.528	-0.177***
vs_pt_2010	0.378	0.524	-0.146***
vs_pt_2014	0.310	0.507	-0.197***

Note: N = 1482 (municipalities for which Twitter data is available, after data cleaning). All variables are aggregated at the municipality level. Column "Lost_m" refers to the municipalities where Bolsonaro lost the 2018 election, whereas column "Won_m" refers to where he won. The third column reports the statistical difference between the respective means. Variables "cellphone," "computer," and "internet" are the proportion of households reported to have such goods in the 2010 Population Census. Variables "no_religion," "catholic," "pentecostal," "black," "indigenous," and "brown" are the proportion of individuals registered to have such demographic characteristics in the 2010 Population Census. Variables "primary" and "tertiary" refers to the population with (at most) primary and tertiary education. Variables "bornhere_mun" and "bornhere_state" refer to the proportion of individuals born in the municipality and state where they answered the 2010 Population Census. Variables "vs_pt_2006," "vs_pt_2010," and "vs_pt_2014" are the proportion of votes obtained by the Workers' Party (Partido dos Trabalhadores, PT) in the $1^{\underline{o}}$ round of the 2006, 2010, and 2014 Presidential Elections, respectively.

Table 3.1 presents descriptive statistics for the municipalities that fall into the treatment and control groups.[19] All demographic variables were extracted from the 2010 Brazilian Population Census. The last three variables are the vote share obtained by the Workers' Party (Partido dos Trabalhadores, PT) in the $1^{\underline{o}}$ round of the 2006, 2010, and 2014 Presidential Elections, respectively. In 2018, Bolsonaro

[19] In Appendix 3.A.3, we present an analogous table for all municipalities. The results in one and other tables do not vary substantially.

defeated a candidate affiliated with the Workers' Party, which explains the negative correlation in votes.

As can be seen, demographic characteristics vary for municipalities in the control and treatment groups, but importantly, the variation is relatively small. For example, regarding the availability of the internet at home, closely related to the presence in social media and Twitter, this difference was 0.04p.p. in 2010. The variable that varies the most is the income per capita. However, as we include Municipality FE in our regressions, these differences are not a threat to identification as long as they are constant over time.

Let us turn to the regression results. Table 3.2 answers this paper's question, *how the 2018 presidential election of Bolsonaro affected online hate speech.* The first column in the table corresponds to the classic difference-in-differences estimation, presented in equation (3.3.1). The second column corresponds to the difference-in-differences model with a continuous treatment variable, i.e., equation (3.3.3). In the two models, we define $Post_t$ as a dummy variable taking a value of one between July 2017 and July 2018 (both included); and a value of zero between January and December 2019 (both included). We drop the period from the election rally to when Bolsonaro took office, as hate speech may behave differently than in regular times. However, in Appendix 3.A.3, we show our results are robust to changes in the definition of $Post_t$.

Column (1) shows the increase in hate speech after the elections that we observe in Figures 3.2 and 3.3 was more pronounced in municipalities where Bolsonaro lost (0.4 p.p. higher). Consistent with this evidence, column (2) shows that the proportion of hate speech decreases as the share of votes for Bolsonaro increases. As the estimate in column (2) comes from a difference-in-differences model with a continuous treatment variable, provided the strong parallel trends assumption, the coefficient is a positively weighted average of the average causal response $ACR(d)$ parameters across doses. Thus, on average, across doses, an increase of 1 p.p. in $VoteShare_m$ decreases hate speech in that municipality by 0.01 p.p.

Focusing on the sign of the estimates in columns (1) and (2), we interpret the results in lines of a beliefs' update mechanism. After receiving the information shock, i.e., the 2018 election result, people could update their beliefs about what type of speeches are socially acceptable. The difference in the election result at the municipality and national levels measures the extent of this update in beliefs. Thus, it is natural to observe that the change in online behavior comes from the individuals who misperceived the social norm before, i.e., those who live in a municipality where Bolsonaro lost. After updating the social norm, they may feel entitled to generate hate speech, even if they live in a municipality with an ex-ante lower level of hate speech.

Table 3.2 Municipality level regressions.

Variables	(1) $Hate_{mt}$	(2) $Hate_{mt}$
$Post_t$ X $Lost_m$	0.004***	
	(0.001)	
$Post_t$ X $VoteShare_m$		-0.010***
		(0.003)
Constant	0.084***	0.087***
	(0.000)	(0.001)
Municipality FE	Yes	Yes
Date FE	Yes	Yes
Municipalities	1,482	1,482
Observations	89,865	89,865
R-squared	0.074	0.073

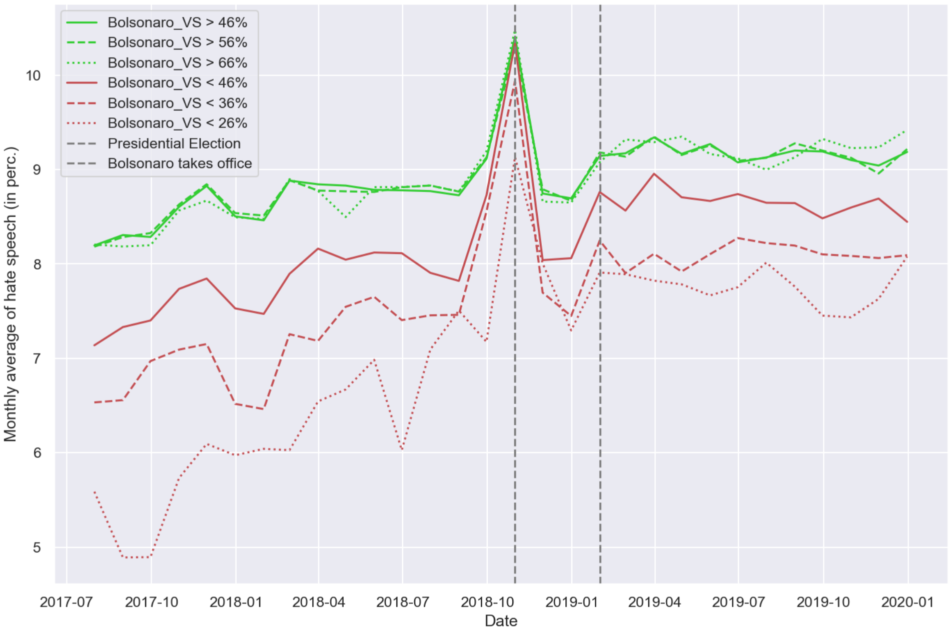
Note: Standard errors in parentheses. $Post_t$ is a dummy variable that takes a value of 0 from July 2017 to July 2018 and a value of 1 from January to December 2019. $Lost_m$ is a dummy variable that takes a value of 1 for the municipalities where Bolsonaro's vote share was lower than 46% and a value of 0 otherwise. *** p<0.01, ** p<0.05, * p<0.1.

Figure 3.4 provides further evidence of this mechanism.[20] This figure is analogous to Figure 3.3, but now splitting hate speech trends between different treatment intensities. For illustrative purposes, hate speech ratios are aggregated at the monthly level. The green lines are the hate speech ratios for the municipalities in which $Lost_m = 0$. The solid, dashed, and dash-dotted lines correspond to municipalities where Bolsonaro got between 46% and 56%, 56% and 66%, and more than 66%, respectively, of the votes in the $1^{\underline{o}}$ round of the 2018 presidential election. On the contrary, the red lines are the monthly proportion of Brazilian tweets classified as hate speech for the municipalities in which $Lost_m = 1$. The solid, dashed, and dash-dotted lines correspond to municipalities where Bolsonaro got between 36% and 46%, 26% and 36%, and less than 26% of the votes, respectively.

There are two relevant observations to this figure. On the one hand, hate speech trends in municipalities where Bolsonaro was popular are similar and relatively stable on time. If individuals living in such municipalities perceived the social norm more accurately even before the elections, their resulting behavior change after elections becomes smaller. On the other hand, hate speech trends in municipalities where Bolsonaro was unpopular were very different in the pre-election period. All the hate speech trends have a positive slope, which is negatively correlated with Bolsonaro's popularity. This negative correlation suggests that the size of the information shock is relevant to explain the extent to which people change their behavior.

[20] An important technical comment on this graph is that our data is unbalanced regarding municipalities where Bolsonaro won and lost, so dashed and, especially, dash-dotted red lines are drawn with a relatively low number of data points. This fact leads us to take this figure's interpretation with some caution.

Figure 3.4 Evolution of hate speech in Brazilian tweets, 2017-2019. Heterogeneity analysis by margins of difference in the 2018 election result.



Note: Hate speech trends are constructed separately for each group of municipalities. "BVS" stands for Bolsonaro's vote share in the $1^{\underline{o}}$ round of the 2018 election.

3.4.2 Individual level

In the previous section, we have shown that the proportion of online hate speech increased after the 2018 presidential election. At the municipality level, this increase is mainly driven by regions where Bolsonaro *lost* the election. As our Twitter data is at the individual level, we can further extend our main analysis and explore *who* is driving this result. In particular, this increase may be driven by (i) users already posting tweets with hate content before the elections, i.e., intensive margin, (ii) users who start posting hate speech tweets after the elections, i.e., extensive margin, or (iii) both.

Throughout this section, we focus on a sub-sample of Twitter users whose tweets are located in *no more than two* different municipalities. For the regression analysis, we restrict our attention to the sub-sample of Twitter users (i) who appear at least one month before and one month after the election and (ii) such that we observe at least 5 tweets per user per month. When a user's tweets are located in multiple municipalities, we assume the information shock she received is a *weighted average* of Bolsonaro's vote share in the corresponding locations.

Figure 3.5 shows that the rise in hate speech results from both the intensive and extensive margins. This figure is constructed using all the Twitter accounts in our sample (after restricting it to the upper bound on the number of locations). Panel a shows how the share of Twitter accounts posting zero hate content becomes smaller after the elections. Specifically, 64.4% of the Twitter users in our sample had never published hate speech content before the 2018 elections, and this number reduced to 62.4% after Bolsonaro was elected president. This reduction is stronger for the sub-sample of Twitter users who post tweets from a municipality where $Lost_m = 1$ - the corresponding percentages are 73.0% in the pre-election period and 69.6% in the post-election period.

Panel b focuses on the intensive margin by zooming in on those Twitter accounts that have posted messages with hate speech at least once. We can see that the distribution of individual hate speech has shifted to the right after the elections. On average, the intensity of hate speech increased by 1.1 p.p. in the post-election period (from 18.2% to 19.3% in the sub-sample of users who posted hate speech content).

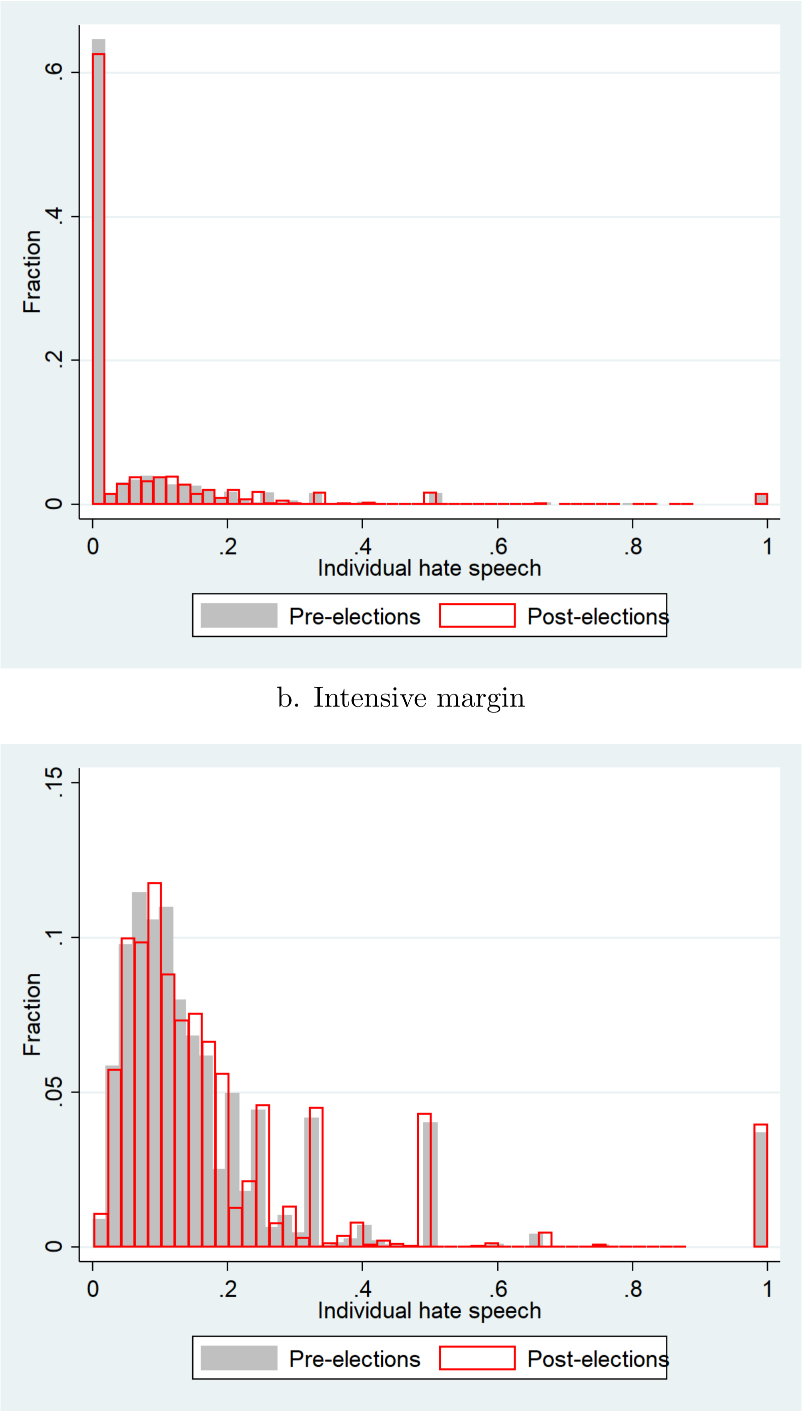
This observation is further confirmed by the estimates in Table 3.3, corresponding to the difference-in-differences models at the individual level, equations (3.3.3) and (3.3.4). We focus on the intensive margin of hate speech, dropping the user accounts that did not publish hateful content during the period under study. Specifically, we delete all Twitter users for whom the proportion of tweets classified as containing hate speech over the period is lower than 5%. Although we lost power in estimating the election's impact on individual hate speech, especially for the continuous treatment differences-in-differences model, the estimates are comparable in sign and magnitude to those previously presented in Table 3.2. In Appendix 3.A.3, we present supplementary regressions, redefining the intensive margin of hate speech and restricting the sub-sample of users according to their online activity.

Table 3.3 Intensive margin of hate speech. Individual regressions.

Variables	(1) $Hate_{imt}$	(2) $Hate_{imt}$
$Post_t$ X $Lost_{im}$	0.003**	
	(0.001)	
$Post_t$ X $VoteShare_{im}$		-0.009
		(0.006)
Constant	0.108***	0.116***
	(0.000)	(0.005)
Individual FE	Yes	Yes
Month FE	Yes	Yes
Individuals	85,494	85,494
Observations	418,616	418,616
R-squared	0.257	0.257

Standard errors in parentheses. Only Twitter users for whom the proportion of tweets classified as containing hate speech over the period is greater than 5%. $Post_t$ is a dummy variable that takes a value of 0 from July 2017 to July 2018 and a value of 1 from January to December 2019. $Lost_m$ is a dummy variable that takes a value of 1 for the municipalities where Bolsonaro's vote share was lower than 46% and a value of 0 otherwise. *** p<0.01, ** p<0.05, * p<0.1.

Figure 3.5 Individual hate speech, pre- and post-elections.



a. Extensive margin

b. Intensive margin

Note: Histograms at the individual level. The pre-election period is between July 2017 and July 2018, whereas the post-election period goes from January to December 2019. Panel a: all the Twitter accounts in the sample. Panel b: only Twitter accounts that posted hate speech content at least once.

3.4.3 Robustness checks

In this section, we check the robustness of the results by relaxing the assumptions we made throughout the paper. For the regressions at the municipality level, we change the variables' definitions and the period under study, among others. For the individual-level regressions, we present results for all Twitter users in the sample, redefine the intensive margin of hate speech, and restrict the sub-sample of users according to their online activity. Appendix 3.A.3 presents the corresponding results, showing that the main results of this paper remain qualitatively unchanged.

3.5 Conclusion

As social media platforms have proliferated, a new public sphere where individuals share ideas has emerged. Among them are the ones related to hate speech, offensive language, and discrimination. Understanding what factors impact the online spread of these harmful speeches is crucial for modern societies, especially regarding social media content moderation. In this line, we provide novel evidence on how political outcomes impact online expressions of hate.

We document that the 2018 election of Bolsonaro in Brazil, a far-right candidate, increased online hate speech. Interestingly, this impact is more pronounced in regions where Bolsonaro was relatively less popular - according to the regression results at the municipality and individual levels. Then, we propose a beliefs update regarding the social acceptability of hate speech as an underlying mechanism.

There are at least three natural extensions of this project, which are left for future research. Firstly, to go deeply into the underlying mechanism proposed in this paper, we can differentiate types of hate speech and analyze their comparative patterns. For that purpose, we plan to develop a machine-learning model to differentiate hate content by its *target*. In the context of this paper, we are particularly interested in the trajectories of hate speech labeled as “homophobia,” “racism,” and “sexism.”

Secondly, an extension of this paper will look at the persistence of information shocks that potentially trigger both a social norms update and the spread of harmful expressions. In our study case, we did not extend the period under analysis as it would require going over the Covid19 pandemic - a completely different type of shock. Lastly, comparing on and offline expressions of hate, especially analyzing their interdependency, is a policy-relevant question; closely related to this paper.

3.A Appendix to Chapter 3

3.A.1 Twitter data

Twitter is an online platform that allows users to publish short messages, of a maximum of 140 characters, on their profiles. In January 2021, Twitter launched an Academic Research product track, which enables researchers to access all v2 endpoints. Notably, the *Twitter Search API v2* gives access to the entire history of public conversations and not only recent tweets. To collect the Twitter data used in this paper, we relied on the *v2 full-archive search endpoint*. We collected tweets using the command line tool and Python library, twarc2 from June 2022 to May 2023. For more information about the academic track on Twitter, follow this link.

The Twitter query we create to download tweets restricts our search to all publicly available (yet undeleted) tweets written in Portuguese, geo-located in Brazil, that are not retweets, and belong to any Monday between July 2017 to December 2019, both included. This query imposes two main assumptions on our tweets' sample. We assume the sample of (i) tweets posted on any Monday and (ii) geo-located tweets are representative samples of the tweets' universe. Figures below present supportive evidence for these assumptions.

Figure 3.6 presents the average number of tweets per day of the week for the period under study. The figure shows that the amount of tweets is quite stable over the weekdays and slightly decreases on weekends. The daily average of tweets is around 305.000. Figure 3.7 shows the daily amount of tweets retrieved by the Twitter query used in this paper but without the restriction of being posted on a Monday. The red dashed line corresponds to the monthly average of tweets.[21] It can be seen that both lines move closer. Furthermore, the monthly trend in Figure 3.7 exhibits higher variation than the average number of tweets per weekday in Figure 3.6, suggesting data for one day per week correctly captures how data behaves.

Lastly, Figures 3.8 and 3.9 compare the trends of geo-located tweets and the universe of tweets that contain a specific word. In all the sub-graphs of the two figures, the red line corresponds to the amount of geo-located tweets, and the blue line is the amount of all tweets multiplied by a *scalability factor*. This factor is the ratio of geo-located tweets over total tweets in the sample for each word, which is between 4% and 8%.

[21] Peaks during June/July 2018 corresponds to dates when Brazil's football team played a match in the 2018 World Cup.

Figure 3.6 Average number of tweets per day in the tweets' sample.

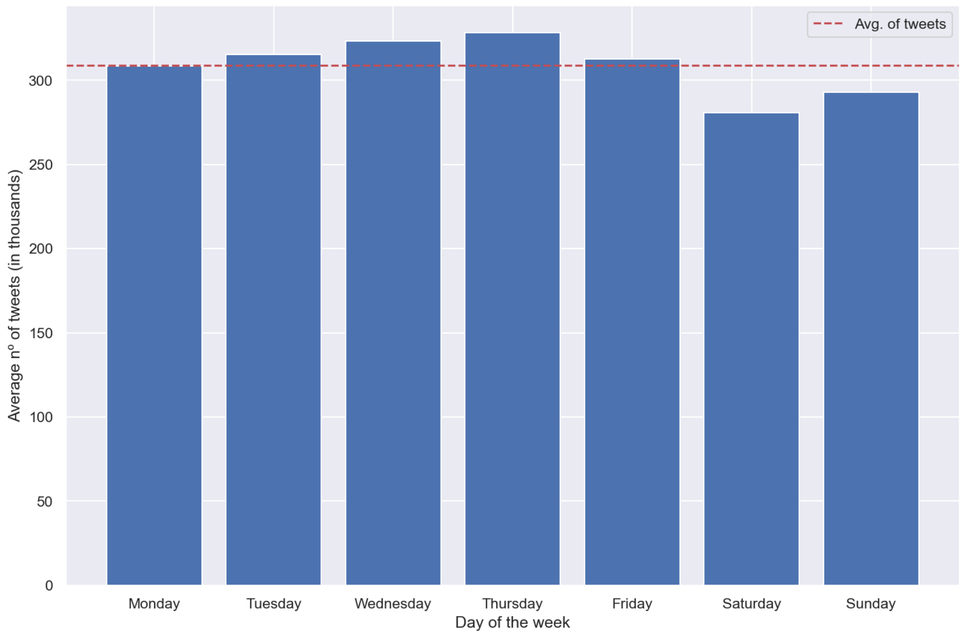
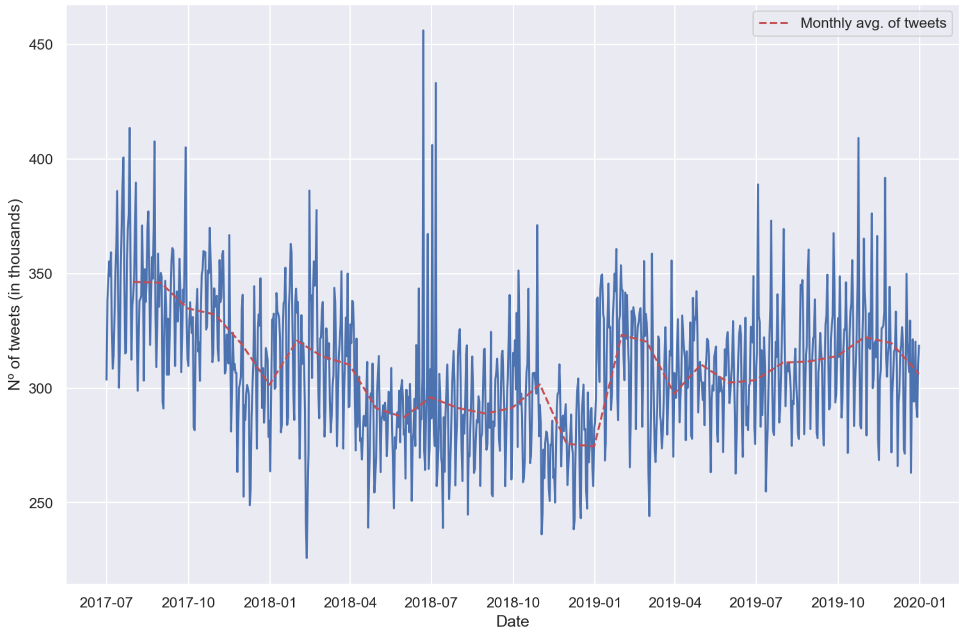


Figure 3.7 Trend of geo-located and total tweets.



In Figure 3.8, the words used are: *"Bolsonaro," "braço"* (arm), *"bom"* (good), *"cão"* (dog), *"cerveja"* (beer), and *"hoje"* (today). In Figure 3.9, we use sensitive words - that may reflect hate speech. Specifically, these words are: *"mariquinha"* (offensive word for a gay man) *"sapatão"* (offensive word for a lesbian), *"nego"* (black), *"preto"* (black), and *"piranha"* and *"putinha"* (offensive words for a woman). As can be seen, both trends behave similarly for each word, suggesting that the subsample of geo-located tweets correctly captures how the universe of tweets behaves. This is especially true for the tweets containing "Bolsonaro."

Figure 3.8 Daily count of tweets retrieved by the Twitter query.

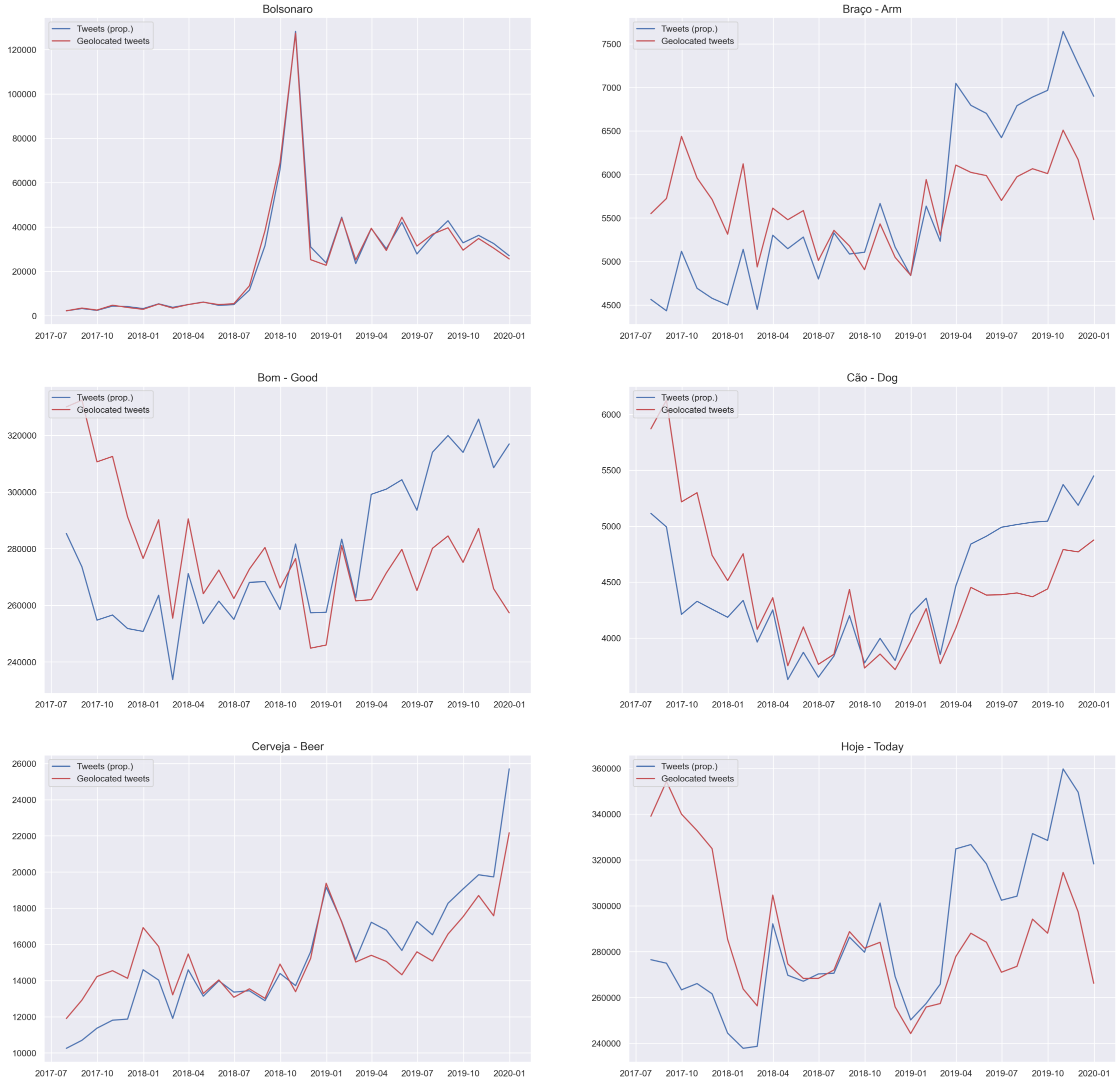
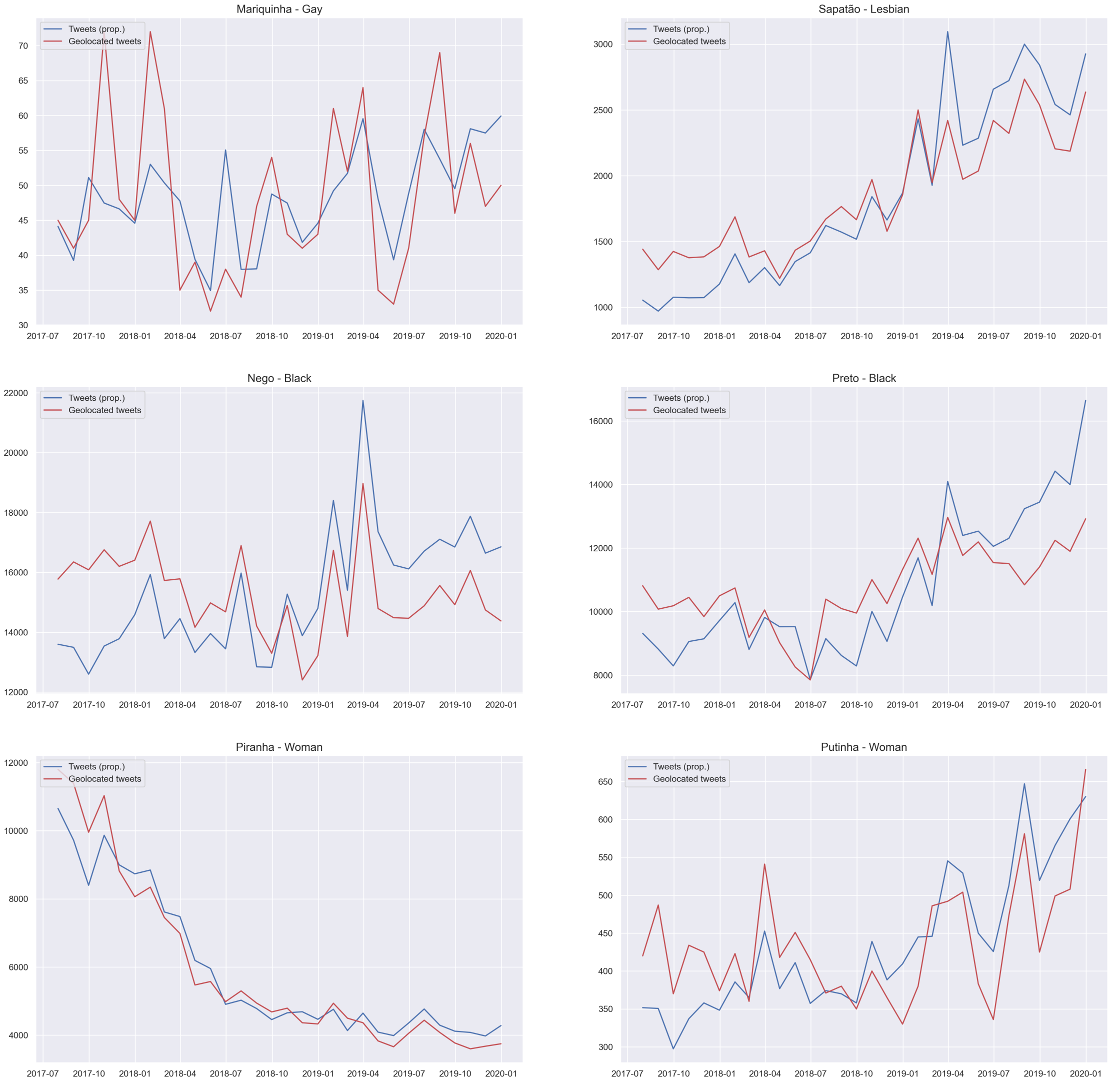


Figure 3.9 Daily count of tweets retrieved by the Twitter query.



3.A.2 Hate speech detection

In this paper, we fine-tune a BERT model on a dataset specific to the hate speech detection task. Fine-tuning is the technique of training a pre-trained model on a suitable dataset for a new task. We use a BERT model in Portuguese, by Souza et al. (2020), and a dataset of tweets in Portuguese, by Fortuna et al. (2019). In the next paragraphs, we describe the resources and procedure.

Model. In this paper, we take *BERTimbau*, a BERT model for Brazilian Portuguese by Souza et al. (2020), as a base model and fine-tune it for the hate speech detection task. Souza et al. (2020) present the model in two sizes: Base (12 layers, 768 hidden dimensions, 12 attention heads, and 110M parameters) and Large (24 layers, 1024 hidden dimensions, 16 attention heads, and 330M parameters). The authors train the models in two tasks: Masked Language Modeling (MLM) and

Sentence Prediction (NSP). The model training is based on the *brWaC corpus* by Wagner Filho et al. (2018), the largest open Portuguese corpus. After training, they evaluate the model in other traditional NLP tasks, namely, Sentence Textual Similarity (STS), Recognizing Textual Entailment (RTE), and Named Entity Recognition (NER). The model improves the state-of-the-art on these tasks, outperforming Multilingual BERT models. The authors made their models publicly available at these Hugging Face links: Base, and Large.

Dataset. We relied on the dataset presented by Fortuna et al. (2019) to fine-tune the BERT model for the hate speech detection task. It is a dataset of tweets in Portuguese collected through Twitter's API, and it comprises 5668 tweets in the period from January to March 2017. The authors provide two annotation schemes for the dataset, binary and hierarchical multiple classifications. For the first classification, three annotators classified every tweet. Each of them had to label the tweet as "hate speech" or "not hate speech," and the authors applied the majority vote to determine the final annotation of each tweet. As a result, 31.5% of the tweets were annotated as "hate speech" on the binary classification dataset. For the hierarchical classification, the authors followed a Rooted Directed Acyclic Graph (DAG) in which "hate speech" is the graph's root. The second level of classes relates to the target of hate, and it comprises: "sexism," "body," "origin," "homophobia," "ideology," "religion," "health," and "lifestyle." As a result, 22% of the tweets were annotated as "hate speech" on the multi-labeled dataset. The authors made their datasets publicly available at this GitHub repository.

Text pre-processing. During text pre-processing, we follow Fortuna et al. (2019) and remove stop-words and punctuation marks using the *NLTK* and *re* Python libraries, respectively. Unlike the authors, we do not remove negative stop-words that may change the statement's meaning. Explicitly, we keep the words: *"mas"* (but), *"nem"* (neither), *"não"* (no), *"sem"* (without), and *"fora"* (out). In addition, we anonymize Twitter mentions as "@user" and links as "URL." We keep "#Hashtags" in the native Twitter format. Finally, we do not transform text to lowercase for consistency with the architecture of Souza et al. (2020) 's BERT model.

Model fine-tuning. We divide the dataset between 80% for training, 10% for validation, and 10% for testing. In NLP applications, the performance of a model in a given task is directly influenced by the characteristics of the training sample. In the case of Fortuna et al. (2019)'s dataset, as in other datasets on hate speech or offensive comments detection, a class imbalance exists. 31.5% of tweets were annotated as "hate speech" under the binary classification, being this minority class. As class imbalance may affect the model's performance, we use a Random Oversampling technique to equalize the number of tweets in the minority and majority classes in the training sample (80% of the tweets). The random oversampling approach randomly adds examples from the minority class to the original training dataset, with replacement.

Training results. The hate-speech BERT model we train attains an overall accuracy of 77% in both the validation and test datasets. Table 3.4 summarizes additional statistics (Precision, Recall, and the F1-score) to characterize the model performance fully.

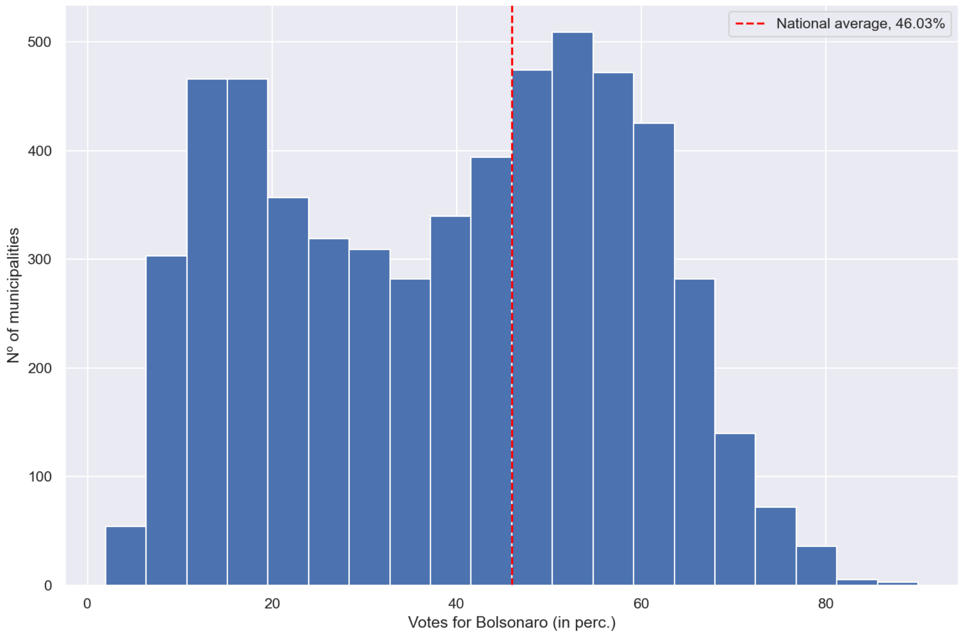
Table 3.4 Training results

Validation sample				
	Precision	Recall	F1	Support
0	0.86	0.76	0.81	365
1	0.64	0.78	0.71	202
W. Avg.	0.79	0.77	0.77	567
Test sample				
	Precision	Recall	F1	Support
0	0.88	0.79	0.83	406
1	0.58	0.72	0.64	161
W. Avg.	0.79	0.77	0.78	567

3.A.3 Tables and Figures

Figure 3.10 Bolsonaro's vote share at the municipality level. 2018 Presidential Election.

a. 1º Round, October 7th.



b. 2º Round, October 28th.

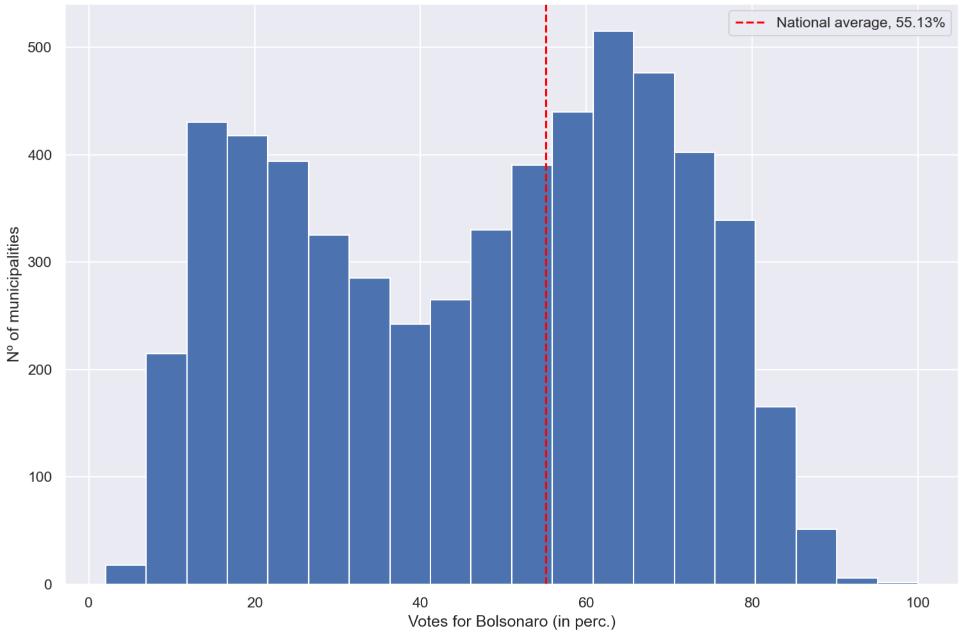
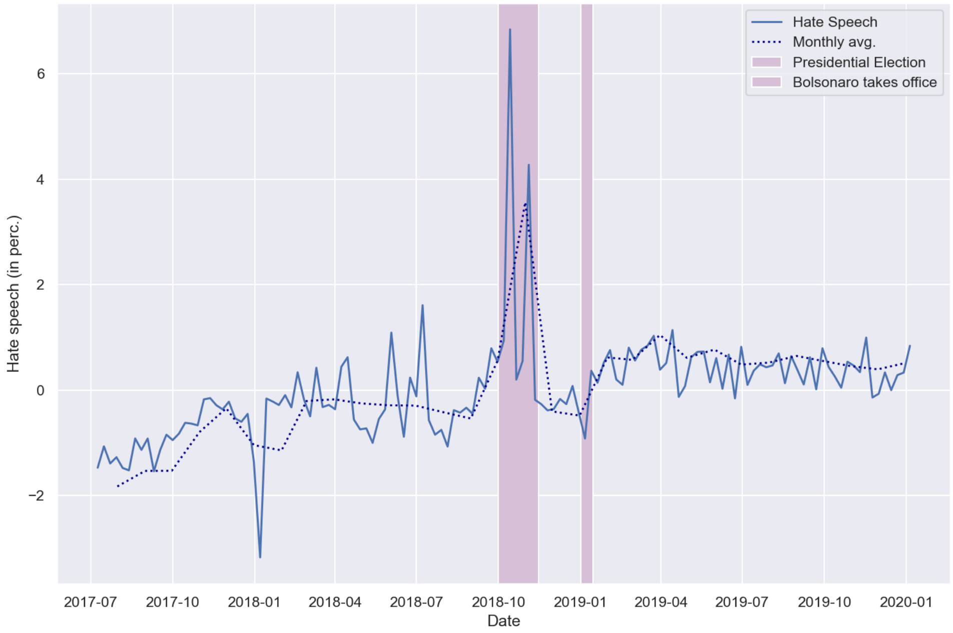
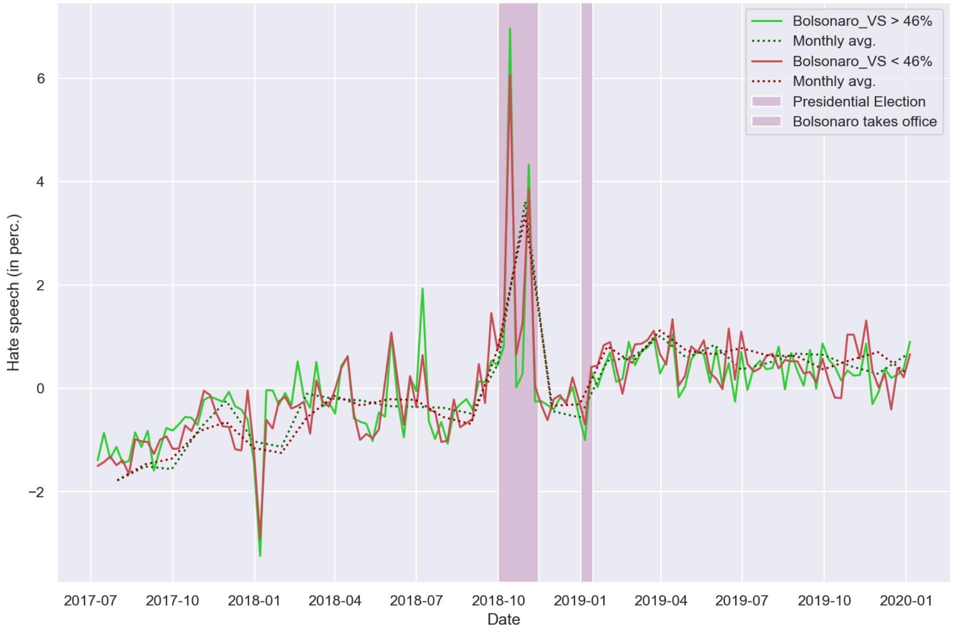


Figure 3.11 Evolution of hate speech in Brazilian tweets, 2017-2019. Standardized variables.

a. All municipalities.

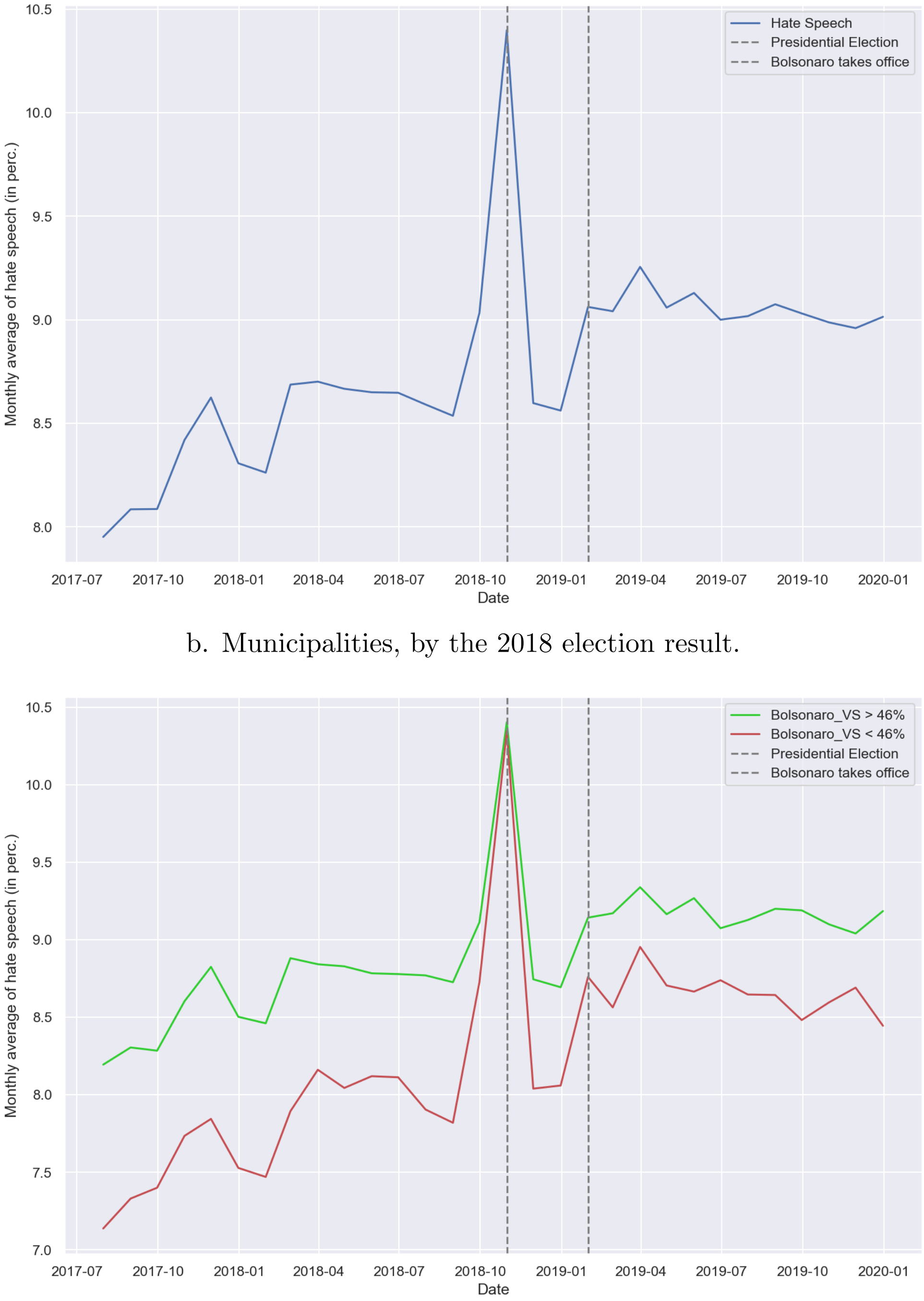


b. Municipalities, by the 2018 election result.



Note: Each hate speech variable was standardized to have zero mean and unit standard deviation.

Figure 3.12 Evolution of hate speech in Brazilian tweets, 2017-2019. Monthly average.



a. All municipalities.

b. Municipalities, by the 2018 election result.

Note: Hate speech variables were aggregated by month.

Table 3.5 Descriptive Statistics, by the 2018 election result.

Variable	Lost_m	Won_m	Difference
urban	0.569	0.735	-0.166***
income_pc	346.8	680.4	-333.6***
cellphone	0.697	0.869	-0.173***
computer	0.140	0.363	-0.223***
internet	0.610	0.703	-0.094***
primary	0.383	0.407	-0.023***
tertiary	0.495	0.425	0.070***
no_religion	0.052	0.052	-0.001
catholic	0.793	0.705	0.087***
pentecostal	0.098	0.126	-0.028***
black	0.075	0.048	0.027***
indigenous	0.009	0.005	0.004***
brown	0.558	0.304	0.255***
born_mun	0.710	0.556	0.154***
born_state	0.742	0.675	0.068***
vs_pt_2006	0.553	0.344	0.210***
vs_pt_2010	0.575	0.391	0.184***
vs_pt_2014	0.585	0.331	0.254***

Note: N = 5570 (municipalities). All variables are aggregated at the municipality level. Column "Lost_m" refers to the municipalities where Bolsonaro lost the 2018 election, whereas column "Won_m" refers to where he won. The third column reports the statistical difference between the respective means. Variables "cellphone," "computer," and "internet" are the proportion of households reported to have such goods in the 2010 Population Census. Variables "no_religion," "catholic," "pentecostal," "black," "indigenous," and "brown" are the proportion of individuals registered to have such demographic characteristics in the 2010 Population Census. Variables "primary" and "tertiary" refers to the population with (at most) primary and tertiary education. Variables "bornhere_mun" and "bornhere_state" refer to the proportion of individuals born in the municipality and state where they answered the 2010 Population Census. Variables "vs_pt_2006," "vs_pt_2010," and "vs_pt_2014" are the proportion of votes obtained by the Workers' Party (Partido dos Trabalhadores, PT) in the 1º round of the 2006, 2010, and 2014 Presidential Elections, respectively.

Regression results at the municipality level.

Table 3.6 Municipality level regressions. Standardized variables.

Variables	(1) $Hate_{mt}$	(2) $Hate_{mt}$
$Post_t$ X $Lost_m$	0.062***	
	(0.016)	
$Post_t$ X $VoteShare_m$		-0.025***
		(0.008)
Constant	-0.019***	-0.010***
	(0.004)	(0.003)
Municipality FE	Yes	Yes
Date FE	Yes	Yes
Municipalities	1,482	1,482
Observations	89,865	89,865
R-squared	0.074	0.073

Note: Standard errors in parentheses. $Post_t$ is a dummy variable that takes a value of 0 from July 2017 to July 2018 and a value of 1 from January to December 2019. $Lost_m$ is a dummy variable that takes a value of 1 for the municipalities where Bolsonaro's vote share was lower than 46% and a value of 0 otherwise. $Hate_{mt}$ and $VoteShare_m$ are standardized to have zero mean and unit standard deviation. *** p<0.01, ** p<0.05, * p<0.1.

Table 3.7 Municipality level regressions. Monthly data.

Variables	(1) $Hate_{mt}$	(2) $Hate_{mt}$
$Post_t$ X $Lost_m$	0.003**	
	(0.001)	
$Post_t$ X $VoteShare_m$		-0.011**
		(0.004)
Constant	0.084***	0.087***
	(0.000)	(0.001)
Municipality FE	Yes	Yes
Date FE	Yes	Yes
Municipalities	1,482	1,482
Observations	27,324	27,324
R-squared	0.144	0.144

Note: Standard errors in parentheses. $Post_t$ is a dummy variable that takes a value of 0 from July 2017 to July 2018 and a value of 1 from January to December 2019. $Lost_m$ is a dummy variable that takes a value of 1 for the municipalities where Bolsonaro's vote share was lower than 46% and a value of 0 otherwise. *** p<0.01, ** p<0.05, * p<0.1.

Table 3.8 Municipality level regressions. Redefining $Post_t$.

Variables	(1) $Hate_{mt}$	(2) $Hate_{mt}$
$Post_t$ X $Lost_m$	0.003***	
	(0.001)	
$Post_t$ X $VoteShare_m$		-0.008***
		(0.003)
Constant	0.084***	0.086***
	(0.000)	(0.001)
Municipality FE	Yes	Yes
Date FE	Yes	Yes
Municipalities	1,482	1,482
Observations	100,375	100,375
R-squared	0.071	0.071

Note: Standard errors in parentheses. $Post_t$ is a dummy variable that takes a value of 0 from July 2017 to August 2018 and a value of 1 from November 2018 to December 2019. $Lost_m$ is a dummy variable that takes a value of 1 for the municipalities where Bolsonaro's vote share was lower than 46% and a value of 0 otherwise. *** p<0.01, ** p<0.05, * p<0.1.

Regression results at the individual level.

Table 3.9 Intensive margin of hate speech. Individual level regressions. Sub-sample of Twitter users, restricted by their activity.

Variables	(1) $Hate_{imt}$	(2) $Hate_{imt}$	(3) $Hate_{imt}$	(4) $Hate_{imt}$
$Post_t$ X $Lost_{im}$	0.002*		0.003**	
	(0.001)		(0.001)	
$Post_t$ X $VoteShare_{im}$		-0.005		-0.005
		(0.006)		(0.006)
Constant	0.107***	0.111***	0.111***	0.125***
	(0.000)	(0.006)	(0.000)	(0.007)
Individual FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Individuals	52,518	52,518	24,180	24,180
Observations	342,732	342,732	210,565	210,565
R-squared	0.260	0.260	0.226	0.226

Standard errors in parentheses. Columns (1)-(2): all Twitter users who posted at least 25 tweets over the period, from which at least 5 are classified as hate speech. Columns (3)-(4): all Twitter users who posted at least 50 tweets over the period, from which at least 10 are classified as hate speech. $Post_t$ is a dummy variable that takes a value of 0 from July 2017 to August 2018 and a value of 1 from November 2018 to December 2019. $Lost_m$ is a dummy variable that takes a value of 1 for the municipalities where Bolsonaro's vote share was lower than 46% and a value of 0 otherwise. *** p<0.01, ** p<0.05, * p<0.1.

Table 3.10 Individual level regressions.

	(1)	(2)
Variables	$Hate_{imt}$	$Hate_{imt}$
$Post_t$ X $Lost_{im}$	0.002	
	(0.001)	
$Post_t$ X $VoteShare_{im}$		-0.005
		(0.005)
Constant	0.090***	0.098***
	(0.000)	(0.004)
Individual FE	Yes	Yes
Month FE	Yes	Yes
Individuals	113,127	113,127
Observations	523,458	523,458
R-squared	0.342	0.342

Standard errors in parentheses. All Twitter users in the sample. $Post_t$ is a dummy variable that takes a value of 0 from July 2017 to August 2018 and a value of 1 from November 2018 to December 2019. $Lost_m$ is a dummy variable that takes a value of 1 for the municipalities where Bolsonaro's vote share was lower than 46% and a value of 0 otherwise. *** p<0.01, ** p<0.05, * p<0.1.

Table 3.11 Individual level regressions. Sub-sample of Twitter users, restricted by their activity.

	(1)	(2)	(3)	(4)
Variables	$Hate_{imt}$	$Hate_{imt}$	$Hate_{imt}$	$Hate_{imt}$
$Post_t$ X $Lost_{im}$	0.002*		0.002**	
	(0.001)		(0.001)	
$Post_t$ X $VoteShare_{im}$		-0.004		-0.005
		(0.005)		(0.005)
Constant	0.091***	0.098***	0.091***	0.097***
	(0.000)	(0.005)	(0.000)	(0.005)
Individual FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Individuals	90,355	90,355	50,329	50,329
Observations	475,402	475,402	357,506	357,506
R-squared	0.315	0.315	0.270	0.270

Standard errors in parentheses. All Twitter users in the sample who posted (i) at least 25 tweets over the period in columns (1)-(2) and (ii) at least 50 tweets in columns (3)-(4). $Post_t$ is a dummy variable that takes a value of 0 from July 2017 to August 2018 and a value of 1 from November 2018 to December 2019. $Lost_m$ is a dummy variable that takes a value of 1 for the municipalities where Bolsonaro's vote share was lower than 46% and a value of 0 otherwise. *** p<0.01, ** p<0.05, * p<0.1.

Bibliography

Acemoglu, D. and Jackson, M. O. (2017). Social norms and the enforcement of laws. *Journal of the European Economic Association*, 15(2):245–295.

Albornoz, F., Bradley, J., and Sonderegger, S. (2022). Updating the social norm: the case of hate crime after the brexit referendum. Technical report, Red Nacional de Investigadores en Economía (RedNIE).

Alesina, A., Giuliano, P., and Nunn, N. (2013). On the origins of gender roles: Women and the plough. *The quarterly journal of economics*, 128(2):469–530.

Ash, E. and Hansen, S. (2023). Text algorithms in economics. *Annual Review of Economics*, 15.

Ballester, C. and Calvó-Armengol, A. (2010). Interactions with hidden complementarities. *Regional Science and Urban Economics*, 40(6):397–406.

Ballester, C., Calvó-Armengol, A., and Zenou, Y. (2006). Who's who in networks. wanted: The key player. *Econometrica*, 74(5):1403–1417.

Belhaj, M., Bramoullé, Y., and Deroïan, F. (2014). Network games under strategic complementarities. *Games and Economic Behavior*, 88:310–319.

Bernheim, B. D. (1994). A theory of conformity. *Journal of political Economy*, 102(5):841–877.

Bhuller, M., Havnes, T., Leuven, E., and Mogstad, M. (2013). Broadband internet: An information superhighway to sex crime? *Review of Economic studies*, 80(4):1237–1266.

Bisin, A. and Verdier, T. (2011). The economics of cultural transmission and socialization. In *Handbook of social economics*, volume 1, pages 339–416. Elsevier.

Boucher, V. (2016). Conformism and self-selection in social networks. *Journal of Public Economics*, 136:30–44.

Bramoullé, Y., Djebbari, H., and Fortin, B. (2009). Identification of peer effects through social networks. *Journal of econometrics*, 150(1):41–55.

Bramoullé, Y., Djebbari, H., and Fortin, B. (2020). Peer effects in networks: A survey. *Annual Review of Economics*, 12:603–629.

Bramoullé, Y., Kranton, R., and D'amours, M. (2014). Strategic interaction and networks. *American Economic Review*, 104(3):898–930.

Bursztyn, L., Cantoni, D., Yang, D. Y., Yuchtman, N., and Zhang, Y. J. (2021). Persistent political engagement: Social interactions and the dynamics of protest movements. *American Economic Review: Insights*, 3(2):233–50.

Bursztyn, L., Ederer, F., Ferman, B., and Yuchtman, N. (2014). Understanding mechanisms underlying peer effects: Evidence from a field experiment on financial decisions. *Econometrica*, 82(4):1273–1301.

Bursztyn, L., Egorov, G., Enikolopov, R., and Petrova, M. (2019). Social media and xenophobia: evidence from russia. Technical report, National Bureau of Economic Research.

Bursztyn, L., Egorov, G., and Fiorin, S. (2020). From extreme to mainstream: The erosion of social norms. *American economic review*, 110(11):3522–3548.

Bursztyn, L., González, A. L., and Yanagizawa-Drott, D. (2018). Misperceived social norms: Female labor force participation in saudi arabia. Technical report, National Bureau of Economic Research.

Cagé, J., Hervé, N., and Mazoyer, B. (2022). Social media and newsroom production decisions.

Callaway, B., Goodman-Bacon, A., and Sant'Anna, P. H. (2021). Difference-in-differences with a continuous treatment. *arXiv preprint arXiv:2107.02637*.

Calvó-Armengol, A. and Ilkılıç, R. (2009). Pairwise-stability and nash equilibria in network formation. *International Journal of Game Theory*, 38(1):51–79.

Cantoni, D., Yang, D. Y., Yuchtman, N., and Zhang, Y. J. (2019). Protests as strategic games: experimental evidence from hong kong's antiauthoritarian movement. *The Quarterly Journal of Economics*, 134(2):1021–1077.

Card, D. and Dahl, G. B. (2011). Family violence and football: The effect of unexpected emotional cues on violent behavior. *The quarterly journal of economics*, 126(1):103–143.

Carr, J., Clifton-Sprigg, J., James, J., and Vujic, S. (2020). Love thy neighbour? brexit and hate crime. Technical report, IZA Discussion Papers.

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., and Flammini, A. (2011). Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, pages 89–96.

Dahl, G. and DellaVigna, S. (2009). Does movie violence increase violent crime? *The Quarterly Journal of Economics*, 124(2):677–734.

De Giorgi, G., Frederiksen, A., and Pistaferri, L. (2020). Consumption network effects. *The Review of Economic Studies*, 87(1):130–163.

De Giorgi, G., Pellizzari, M., and Redaelli, S. (2010). Identification of social interactions through partially overlapping peer groups. *American Economic Journal: Applied Economics*, 2(2):241–75.

De Paula, Á. (2020). Econometric models of network formation. *Annual Review of Economics*, 12:775–799.

DellaVigna, S., Enikolopov, R., Mironova, V., Petrova, M., and Zhuravskaya, E. (2014). Cross-border media and nationalism: Evidence from serbian radio in croatia. *American Economic Journal: Applied Economics*, 6(3):103–132.

DellaVigna, S., List, J. A., Malmendier, U., and Rao, G. (2016). Voting to tell others. *The Review of Economic Studies*, 84(1):143–181.

Echenique, F. (2007). Finding all equilibria in games of strategic complements. *Journal of Economic Theory*, 135(1):514–532.

Edmond, C. (2013). Information manipulation, coordination, and regime change. *Review of Economic studies*, 80(4):1422–1458.

Enikolopov, R., Makarin, A., and Petrova, M. (2020). Social media and protest participation: Evidence from russia. *Econometrica*, 88(4):1479–1514.

Fernandez, R. (2007). Women, work, and culture. *Journal of the European Economic Association*, 5(2-3):305–332.

Fortuna, P., da Silva, J. R., Wanner, L., Nunes, S., et al. (2019). A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online*, pages 94–104.

Genicot, G. (2022). Tolerance and compromise in social networks. *Journal of Political Economy*, 130(1):94–120.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–74.

Gentzkow, M. and Shapiro, J. M. (2011). Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4):1799–1839.

Gerber, A. S., Green, D. P., and Larimer, C. W. (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *American political Science review*, 102(1):33–48.

Giuliano, P. (2007). Living arrangements in western europe: Does cultural origin matter? *Journal of the European Economic Association*, 5(5):927–952.

Goldsmith-Pinkham, P. and Imbens, G. W. (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics*, 31(3):253–264.

González, F. (2020). Collective action in networks: Evidence from the chilean student movement. *Journal of Public Economics*, 188:104220.

Gulesci, S., Jindani, S., La Ferrara, E., Smerdon, D., Sulaiman, M., and Young, H. (2021). A stepping stone approach to understanding harmful norms.

Hager, A., Hensel, L., Hermle, J., and Roth, C. (2023). Political activists as free riders: Evidence from a natural field experiment. *The Economic Journal*, 133(653):2068–2084.

Halberstam, Y. and Knight, B. (2016). Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter. *Journal of public economics*, 143:73–88.

Hsieh, C.-S. and Lee, L. F. (2016). A social interactions model with endogenous friendship formation and selectivity. *Journal of Applied Econometrics*, 31(2):301–319.

Hsieh, C.-S., Lee, L.-F., and Boucher, V. (2020). Specification and estimation of network formation and network interaction models with the exponential probability distribution. *Quantitative economics*, 11(4):1349–1390.

Ivandic, R., Kirchmaier, T., and Machin, S. J. (2019). Jihadi attacks, media and local hate crime.

Jackson, M. O. and Watts, A. (2001). The existence of pairwise stable networks. *Seoul journal of economics*, 14.

Jackson, M. O. and Wolinsky, A. (1996). A strategic model of social and economic networks. *Journal of economic theory*, 71(1):44–74.

Jackson, M. O. and Zenou, Y. (2015). Games on networks. In *Handbook of game theory with economic applications*, volume 4, pages 95–163. Elsevier.

Jackson, S. J., Bailey, M., and Welles, B. F. (2020). *# HashtagActivism: Networks of race and gender justice.* Mit Press.

Larson, J. M., Nagler, J., Ronen, J., and Tucker, J. A. (2019). Social networks and protest participation: Evidence from 130 million twitter users. *American Journal of Political Science*, 63(3):690–705.

Levin, J. (2003). Supermodular games. *Lectures Notes, Department of Economics, Stanford University.*

Levy, G. and Razin, R. (2019). Echo chambers and their effects on economic and political outcomes. *Annual Review of Economics*, 11:303–328.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542.

Milgrom, P. and Roberts, J. (1990). Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica: Journal of the Econometric Society*, pages 1255–1277.

Moretti, E. (2011). Social learning and peer effects in consumption: Evidence from movie sales. *The Review of Economic Studies*, 78(1):356–393.

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

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

Nicoletti, C., Salvanes, K. G., and Tominey, E. (2018). The family peer effect on mothers' labor supply. *American Economic Journal: Applied Economics*, 10(3):206–34.

Olson, M. (2009). *The logic of collective action*, volume 124. Harvard University Press.

Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of economic perspectives*, 14(3):137–158.

Passarelli, F. and Tabellini, G. (2017). Emotions and political unrest. *Journal of Political Economy*, 125(3):903–946.

Patacchini, E., Rainone, E., and Zenou, Y. (2017). Heterogeneous peer effects in education. *Journal of Economic Behavior & Organization*, 134:190–227.

Patacchini, E. and Zenou, Y. (2012). Juvenile delinquency and conformism. *The Journal of Law, Economics, & Organization*, 28(1):1–31.

Perez-Truglia, R. and Cruces, G. (2017). Partisan interactions: Evidence from a field experiment in the united states. *Journal of Political Economy*, 125(4):1208–1243.

Robinson, J. (1951). An iterative method of solving a game. *Annals of mathematics*, pages 296–301.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.

Topkis, D. M. (1979). Equilibrium points in nonzero-sum n-person submodular games. *Siam Journal on control and optimization*, 17(6):773–787.

Ushchev, P. and Zenou, Y. (2020). Social norms in networks. *Journal of Economic Theory*, 185:104969.

Vives, X. (1990). Nash equilibrium with strategic complementarities. *Journal of Mathematical Economics*, 19(3):305–321.

Wagner Filho, J. A., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brwac corpus: a new open resource for brazilian portuguese. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Yanagizawa-Drott, D. (2014). Propaganda and conflict: Evidence from the rwandan genocide. *The Quarterly Journal of Economics*, 129(4):1947–1994.

Young, H. P. (2015). The evolution of social norms. *economics*, 7(1):359–387.

Zhuravskaya, E., Petrova, M., and Enikolopov, R. (2020). Political effects of the internet and social media. *Annual review of economics*, 12:415–438.