The Lifecycle of Protests in the Digital Age

Pierre C. Boyer Germain Gauthier Yves Le Yaouanq

Vincent Rollet Benoît Schmutz-Bloch

January 2026

We propose a theory of the emergence, size, intensity, and duration of modern protest movements. Moderates and radicals are both needed to sustain large coalitions, but when radicals resort to violence, they drive moderates away. Social media, by lowering the cost of mobilization, amplifies this tension: it reveals both the potential for protest and the proportion of radicals among protesters — sparking a mass movement while enabling radicals to coordinate on violent action that precipitates the movement's demise. We illustrate this phenomenon with the 2018 French Yellow Vests uprising. Online mobilization initially helped organize large, peaceful protests, but these protests triggered a second wave of more radical online activity. We show that half of the movement's subsequent radicalization online occurred through the departure of moderates, driven by their exposure to radical content.

Keywords: Protests; Learning traps; Crowding-out; Violence; Social media; NLP.
JEL Codes: D72, D74, L82, Z13.

Boyer: CREST, École Polytechnique, Institut Polytechnique de Paris, France (pierre.boyer@polytechnique.edu); **Gauthier**: Bocconi University, Italy (germain.gauthier@unibocconi.it); **Le Yaouanq**: CREST, École Polytechnique, Institut Polytechnique de Paris, France (yves.le-yaouanq@polytechnique.edu); **Rollet**: MIT, USA (vrollet@mit.edu); **Schmutz-Bloch**: CREST, École Polytechnique, Institut Polytechnique de Paris, France (benoit.schmutz@polytechnique.edu). The authors thank Luca Braghieri, Micael Castanheira, Thomas Delemotte, Allan Drazen, Georgy Egorov, Sophie Hatte, Emeric Henry, Matthew Jackson, David Levine, Clément Malgouyres, Matías Núñez, Paula Onuchic, Harry Pei, Vincent Pons, Mehdi Shadmehr, Clémence Tricaud, Ekaterina Zhuravskaya and Galina Zudenkova, as well as many seminar and conference participants for their comments. Jonathan Garson provided great research assistance. The authors gratefully acknowledge the Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047), ANR-19-CE41-0011-01, ANR-20-CE41-0013-01 and the Chaire Professorale Jean Marjoulet for financial support, the CASD (Centre d'Accès Sécurisé aux Données) and INSEE for the access to French administrative data, and Change.org for sharing their anonymized data.

We must not allow our creative protest to degenerate into physical violence.
Martin Luther King — August 28th, 1963
I have a Dream

1 Introduction

Every year, thousands of protest movements break out around the world (Cantoni, Kao, Yang and Yuchtman, 2024). Some last a few days, others months or even years. Some stay local, others spread to subcontinents. Some are mostly peaceful, others are violent. Last but not least, these movements are only the tip of the unrest iceberg: many others are stillborn. In this paper, we study the lifecycle of modern protest movements — their emergence, size, intensity, and eventual decline — through the lens of a simple political tension. Moderate and radical protesters are both needed to sustain large coalitions, but these groups are uneasy allies. When radicals resort to violence, they can drive moderates away and ultimately undermine the movement. Low-cost mobilization technologies such as social media exacerbate this tension. By reducing participation costs and rapidly aggregating information, social media reveals both the movement's mobilization potential and the proportion of radicals among participants. This can spark mass, initially peaceful protests, but it also helps radical factions coordinate, increasing the likelihood of subsequent escalation and fragmentation. We illustrate this phenomenon by combining online and offline protest data for the Yellow Vest movement, one of the most notable episodes of social unrest in recent French history.

We follow the tradition of modeling protests as a game in which payoffs depend on the total number of players who choose to mobilize. We incorporate heterogeneous protest intensity, as mobilization can be either peaceful or violent, and heterogeneous preferences among protesters, with passive types (who never participate), moderates (who might participate peacefully or abstain), and radicals (who might choose among the three possible actions). The shares of the different types in the population are initially unknown. We assume strategic complementarity within types but strategic substitutability across types, as moderates prefer participating when others are peaceful rather than violent. Protests are thus characterized by two margins, extensive (size) and intensive (level of violence). We analyze the equilibria of this game, and show how the type of protest that emerges from this strategic interaction depends on players' beliefs about the distribution of preferences in the population.

Next, we analyze a dynamic extension of the game in which players update their beliefs about the population's preferences based on past protests. Some equilibria imply

an identification problem. This creates the possibility of learning traps, in which players remain pessimistic about the shares of moderates and/or radicals even with an infinite number of periods, and the long-run equilibrium differs from the equilibrium that would be reached if the population's characteristics were known (Fudenberg and Levine, 1993). These traps can affect the extensive margin of the protest, its intensive margin, or both margins simultaneously. Armed with this concept, we explain why the impact of reduced participation costs on the long-term viability of "large" protests (those involving moderates) is nuanced and not invariably favorable. While low costs may overcome learning barriers at the extensive margin, they may also result in a smaller, more violent movement by helping protesters learn about a higher share of radicals.

We apply these observations to the interplay between street protests and social media, modeled as a lower-cost mobilization technology available before each protest episode. In the short run, social media can foster large and peaceful street protests by revealing that discontent in the population is greater than expected. However, the story does not end here. If these first street protests succeed, they will reveal additional encouraging information about mobilization potential, which may trigger a surge in subsequent online mobilization. These two predictions are particularly amenable to empirical analysis using exogenous variations in the cost of (online, then offline) protests. After this initial *crowding-in* phase where both margins of participation are activated online, several trajectories are possible. If the new wave of online activity indicates a large share of radical players, they will coordinate on violent action and might eventually *crowd out* moderates, both online and offline. This *crowd-in-then-crowd-out* sequence can arise without any bias in social media favoring radical content, but such a bias may compound the risk of crowding out by unduly scaring moderates away. Conversely, state-mandated internet shutdowns during social unrest may artificially prolong the life of a large protest.

In the second part of the paper, we analyze the trajectory of the Yellow Vest movement in light of our conceptual framework. This movement shared many characteristics with concurrent protest movements around the world.[1] Sparked by an online petition

[1]For example, Shultziner and Kornblit (2020) argue that the Yellow Vest movement is quite similar to the Occupy movements in Spain, Israel, Ireland, and the United States in terms of origins (economic issues and relative deprivation), organization (decentralized and deliberately leaderless), and tactics (nationwide occupation of public spaces). It also bears a striking resemblance to the 2013 protests in Brazil, which were initially organized against transportation fare hikes but grew to include other issues such as government corruption and police brutality (Winters and Weitz-Shapiro, 2014). The online-offline dynamics we study have many similarities to the events that unfolded in the United States after the 2020 presidential election. These events included the swift organization

against high gas prices and with strong bipartisan appeal, it used social media (primarily Facebook) to successfully organize hundreds of roadblocks across the country on November 17, 2018 (hereafter 11/17). After this first day of widespread and mostly peaceful protests, the movement remained very active online. At the same time, however, street protests quickly became more violent, drew fewer participants, and polls showed that the movement had lost popular support. To study this movement, we combine geolocated data on street protests, Facebook groups, and petition signatures with textual analysis of a panel of discussions on Facebook pages.

We start by documenting the movement's heavy reliance on social media in its early days, using spatial analysis at the municipality level. Like previous research in other settings, we first show that early online activity led to more roadblocks on 11/17. We then describe a lesser-known phenomenon in the literature: these roadblocks triggered a second wave of online activity in the weeks that followed. According to our model, this rebound effect corresponds to the last stage of crowding-in, whereby the 11/17 protests helped spread information about the popularity of the Yellow Vests, which increased the size of subsequent online mobilization. Consistent with our theoretical predictions, we establish both directions of this feedback loop with two different instrumental variable strategies based on exogenous variation in the cost of online and offline mobilization: the progressive deployment of the 4G network and the spatial distribution of highway tolls, which were heavily targeted by protesters as a symbol of car-related expenses.

Despite this online-offline feedback loop, however, protests quickly subsided after 11/17. To understand the movement's decline, we follow our theoretical framework and examine the relationship between the size of protests and their violence. In the absence of panel data on street protesters, we leverage another dataset of discussions on Yellow Vest Facebook pages, for which we can track individual protesters' comments over time. Using text-as-data techniques (Gentzkow, Kelly and Taddy, 2019; Ash and Hansen, 2023), we analyze the radicalization process of a large group of discussants whose discussions became increasingly radical over time. We first use this dataset to illustrate the Bayesian updating process about the share of radical discussants and show that the many pages created in the immediate aftermath of 11/17 led to a quick upward reevaluation of this share, as predicted by the *crowd-in-then-crowd-out* sequence.

We then exploit the panel dimension of our dataset to decompose the radicalization process over the following months into an extensive margin (changes in the composition

of "Stop the Steal" rallies in late 2020 and the Capitol riots on January 6, 2021. The latter were so severe that they raised concerns about a potential rebellion and arguably brought an end to the movement.

of the population of discussants) and an intensive margin (an increase in the tendency to post radical messages at the individual level). According to our estimates, both margins played almost equally important roles, although the effect of the extensive margin was slightly delayed relative to that of the intensive margin, consistent with a potential crowding-out of moderate discussants by more radical ones at the aggregate level. Finally, we use this empirical framework to provide direct evidence of the crowding out of moderates, who left Facebook pages where discussions had become more radical. This effect is quantitatively important and is robust to controlling for the sorting of discussants across pages and for page-by-period unobservable characteristics.

Relationship to the literature. Our first contribution is to propose a novel model of protest dynamics. The framework we propose has four main features.

First, we conceptualize protests as a coordination game, a standard feature of the literature on collective action (Granovetter, 1978). An important element we add to this literature is the explicit modeling of an intensive margin and of a strategic interaction between different types of protesters.[2] Some of these interactions feature strategic substitutability, allowing for a richer taxonomy of protests relative to the literature, which has so far focused on the case of strategic complements.[3] Empirically, some strategic substitutability is found in the studies by Cantoni, Yang, Yuchtman and Zhang (2019) and Hager, Hensel, Hermle and Roth (2022), who both provide experimental evidence that an upward shift in beliefs about turnout can depress participation. In our framework, substitutability arises if moderates interpret this information as indicative of a large mobilization of radicals. While this is unlikely to be the case in the study by Cantoni et al. (2019), where all subjects are university students, this mechanism is more plausible in the study by Hager et al. (2022), where substitutability is found among supporters of the AfD, a German far right movement.[4]

Second, protesters are imperfectly informed about the preferences of their peers, and learn about it by observing data from past protests. The idea that protesting decisions

[2] A distinct strand of the literature studies the strategic interaction between protesters and the government's response (e.g. Lohmann, 1993; Battaglini, 2017; Morris and Shadmehr, 2023, 2024). Another body of research examines how rebel groups choose between violent or peaceful tactics when managing public opinion (Bueno de Mesquita, 2013; Yao, 2024).

[3] An exception is the paper by Steinert-Threlkeld, Chan and Joo (2022), who provide evidence of crowding out as a result of violent protests.

[4] In the same study, Hager et al. (2022) also find that the treatment effect works in the opposite direction (strategic complementarity) for left-leaning supporters of a counter-protest.

are affected by strategic uncertainty has many precedents, most notably in the literature on global games (Morris and Shin, 1998; Angeletos, Hellwig and Pavan, 2007).[5] In this literature, each individual receives a noisy signal about the strength of the regime. We show that rich dynamics arise even when all players share the same belief about the preferences of the population. We also complement this literature by analyzing the long-run relationship between protesters' beliefs and actions. To do so, we borrow tools from the literature on active learning in games (Fudenberg and Levine, 1993).

Third, we do not only study the birth and size of protests but also their intensity and persistence. We show how the strategic interaction between moderates and radicals can trigger an initial movement of increasing participation followed by a sharp decline (crowd-in-then-crowd-out), a pattern we document empirically in the case of the Yellow Vest movement. A similar dynamic arises in the models by Correa (2025) and Enikolopov, Makarin, Petrova and Polishchuk (2020b), but for different reasons. In Correa (2025), participants drop out gradually to receive reputational rewards contingent on the duration of their participation in the movement. In Enikolopov et al. (2020b), participation is driven by signaling motives and declines over time as the reputational payoff of an extra round of mobilization decreases. Gieczewski and Kocak (2024) study another type of crowding out due to intertemporal substitution in protests. Bursztyn, Cantoni, Yang, Yuchtman and Zhang (2021) study the roots of persistent mobilization empirically. They show that incentives to attend a protest once have dynamic consequences if a significant share of the protester's social network also turns out.

Fourth, we explicitly study the causal effect of social media by assuming that its main role is to facilitate learning about the population's preferences. The closest existing model is that of Barbera and Jackson (2020), who study how the shape of social interactions (prior beliefs, homophily, number of contacts) influences the likelihood of a revolution. An important difference is that Barbera and Jackson (2020) view online political activity as cheap talk (hence inconsequential), while we model it as a costly (hence informative) form of political participation. This view, which is supported by our empirical analysis, allows us to make predictions about the dynamics of protests with and without social media.

We also contribute to the study of the interaction between online and offline forms of protest. A large empirical literature has studied the effect of social media on the emer-

[5] See also Shadmehr and Bernhardt (2011); Kricheli, Livne and Magaloni (2011); Little (2016, 2017). Some papers study information revelation in a different direction, from opinion leaders to followers (e.g. Loeper, Steiner and Stewart, 2014).

gence of protest movements, with most studies finding a positive effect.[6] Conceptually, social media might serve two purposes: aggregating information about the population's preferences, and the concrete planning of protests (e.g., choosing the location).[7] Little (2016) models both channels and shows that the former effect might be negative if the unpopularity of the regime is not as strong as expected. While our model focuses on information aggregation, with social media acting as a petition (Battaglini, Morton and Patacchini, 2020), our empirical section provides a direct illustration of this dual function of social media using data from both a virtual forum (Facebook) and a counting device (Change.org). We also show, using two different methods (high-frequency time series and an IV approach), that online-offline interactions may extend beyond the initial stage and therefore nurture a positive feedback loop that can help protest movements persist and grow, in line with recent evidence on the 15M movement in Spain (Casanueva, 2025).

Finally, we discuss how social media can contribute to the premature end of protest movements. Protests ignited online have been shown to have weaker links between participants and thus fizzle out quickly (Tufekci, 2017). We offer a complementary mechanism based on these protests' higher propensity to radicalize, in line with recent evidence on Twitter in the US (Gylfason, 2025). The main explanation for the effect of social media on user radicalization appeals to built-in biases in social media technology, even though the exact mechanisms remain debated (see, e.g., Ross-Arguedas, Robertson, Fletcher and Nielsen, 2022).[8] We contribute to this debate by introducing a radicalization process that does not rely on such biases and showing why an algorithmic bias towards violent discussions may prevent large protests due to its contradictory effects on the different factions behind the movement. We also add to this literature by proposing several empirical methods to measure radicalization that take advantage of the structure and content of social media data.

[6]See Zhuravskaya, Petrova and Enikolopov (2020) and Aridor, Jiménez-Durán, Levy and Song (2024) for reviews. Specific movements and social media are studied by Acemoglu, Hassan and Tahoun (2018), Larson, Nagler, Ronen and Tucker (2019), Enikolopov, Makarin and Petrova (2020a), or Fergusson and Molina (2021), among others. Other studies focus on different outcomes, such as hate crimes (Bursztyn, Egorov, Enikolopov and Petrova, 2024) or voting behavior (Fujiwara, Muller and Schwarz, 2024).

[7]Beyond information and coordination motives, Enikolopov et al. (2020b) show that large online movements may magnify the reputational incentives to participate offline.

[8]For example, Levy (2021) shows that Facebook's algorithm is less likely to expose users to posts from news outlets with opposing views, which would reduce their negative attitudes towards the opposing political party. Conversely, Bail, Argyle, Brown, Bumpus, Chen, Hunzaker, Lee, Mann, Merhout and Volfovsky (2018) find that Republicans express more conservative views after being exposed to liberal Twitter bots.

The remainder of the paper is organized as follows. Section 2 presents our theoretical framework. We provide empirical evidence of a crowd-in-then-crowd-out sequence on the Yellow Vest Movement in Section 3. Section 4 concludes. Formal proofs and other details about our application are relegated to the Appendix.

2 Conceptual framework

We present a dynamic model of political participation based on strategic uncertainty and information revelation. Strategic interactions imply that the protest dynamics depends on participation costs in non-trivial ways.

2.1 The protest game

Our framework involves repeated protest participation decisions. We start by describing and analyzing the stage game of protests. We are particularly interested in the conditions compatible with the existence of a "large" protest involving different groups of protesters.

2.1.1 Framework

Preferences. We consider a population of agents of mass one. Each agent is characterized by a type $\theta \in \mathbb{R}_+$ that measures the willingness to participate in the protest movement.[9]

Participation decisions take three possible values: abstaining ($a = \mathbf{A}$), participating in a peaceful manner ($a = \mathbf{P}$), and participating in a violent manner ($a = \mathbf{V}$).[10] The utility from not participating is normalized to zero. The utility from protesting depends on seven parameters θ, v, α, β, γ, $\underline{c}$, $\overline{c}$: θ measures the individual-level propensity to protest, v measures the benefit from violent action, α, β, γ measure externalities from other protesters, and $\underline{c}$ and $\overline{c} > \underline{c} > 0$ measure the direct cost of peaceful and violent

[9] This model is consistent with an interpretation of θ as reflecting a protester's expressive concern, or their desire to trigger a policy change. Types do not change over time, consistently with empirical evidence provided by Gethin and Pons (2024) showing that recent protests in the US had limited effect on political attitudes.

[10] Protesters do not make decisions with the purpose of conveying (or collecting) information (unlike, e.g., Bueno de Mesquita, 2010). Indeed, in our model the information is publicly available to everyone (not just to protesters), and there is no scope for costly political participation for the purpose of information provision, as every individual has a negligible impact on aggregate information.

protests, respectively. The payoffs to participating depend on the mass of individuals selecting either type of action: An individual i of type θ_i who plays $a_i = \mathbf{P}$ reaps a payoff equal to

$$U_i[a_i = \mathbf{P}, \{a_j\}] = \theta_i + \alpha \mathbb{E}\mathbb{1}_{a_j=\mathbf{P}} - \beta \mathbb{E}\mathbb{1}_{a_j=\mathbf{V}} - \underline{c} \tag{1}$$

while the same individual playing $a_i = \mathbf{V}$ receives

$$U_i[a_i = \mathbf{V}, \{a_j\}] = (v+1)\theta_i + \alpha \mathbb{E}\mathbb{1}_{a_j=\mathbf{P}} + \gamma \mathbb{E}\mathbb{1}_{a_j=\mathbf{V}} - \overline{c}. \tag{2}$$

Thus, the utility of protesting depends on the intrinsic willingness-to-participate θ (net of the cost), on the type of protest (peaceful or violent), and on the number of participants resorting to either action. Complementarities can reflect the fact that protesting is more useful, more meaningful, or less risky when done in a large crowd.

We assume $\gamma > \alpha > 0, \beta > 0$ and $v > 0$: all interdependencies take the form of a strategic complementarity, except that violent action discourages peaceful protests; besides, complementarities are stronger for violent than for peaceful actions.[11] In addition, more extreme types (higher θ) have a greater gain from choosing violence.

Types and uncertainty. Agents' preferences are heterogeneous. A fraction $1 - \mu$ is *passive* ($\theta = \theta_P \approx -\infty$) and plays $a_P = \mathbf{A}$. Among the remaining, potentially active citizens, a fraction $1 - \lambda$ is *moderate* ($\theta = \theta_M$), while the remaining share λ is *radical* ($\theta = \theta_R > \theta_M$). We further restrict the analysis by assuming that moderates never engage in violent action ($a_M \in \{\mathbf{A}, \mathbf{P}\}$). There are thus two types of movements: "small" protests in which only radicals participate, and "large" protests in which both types are active.

The parameters λ and μ are uncertain. In the dynamic version of the game, from subsection 2.2 onward, players use information about past protests to update their beliefs about λ and μ. We assume that protesters do not make any inference about (λ, μ) from the realization of their own type, so all groups share a common belief. In the stage game, we capture this belief via the expectations $\mathbb{E}[\lambda\mu]$ and $\mathbb{E}[\mu]$ of the share of radical and active individuals, respectively.

Solution concept. We look for pure-strategy Nash equilibria of the stage game where each agent best responds to others' participation decisions given their beliefs about λ

[11] Our main predictions also hold when $\beta < 0$ and $|\beta| < \alpha$, i.e., when peaceful protesters value violent protesters positively, but less so than peaceful ones.

and μ. The stage game admits multiple equilibria for many parameter values. To limit the number of cases to consider, we impose some equilibrium refinements.

Our solution concept refines Nash equilibrium by requesting that, in equilibrium, no group of players (moderates or radicals) can strictly benefit from collectively coordinating on a different action.[12] When multiple equilibria survive this refinement and one of them Pareto-dominates the others, we select it. Last, in knife-edge cases when multiple equilibria persist due to one coalition (or both) being indifferent between two actions, we break ties to obtain a unique equilibrium. Details are in Appendix A.1.

This solution concept rules out inefficient equilibria that result from standard coordination frictions inside political factions. This allows us to focus on the type of inefficiencies at the core of this paper: those that arise because players are imperfectly informed of each other's preferences.

2.1.2 Equilibrium

An equilibrium is described by a pair (a_M, a_R), where $a_M \in \{\mathbf{A}, \mathbf{P}\}$ is the strategy of the moderates and $a_R \in \{\mathbf{A}, \mathbf{P}, \mathbf{V}\}$ is that of the radicals. We interpret the five possible equilibria as follows: $(\mathbf{A}, \mathbf{A})$ corresponds to inaction, or *Rest*; $(\mathbf{A}, \mathbf{P})$ describes a *Routine* situation in which both types choose their default action; if moderates join radicals, they form a *Rally* — equilibrium $(\mathbf{P}, \mathbf{P})$. Radicals, however, may choose to protest violently. If they do so without the support of moderates, they lead a *Riot* — equilibrium $(\mathbf{A}, \mathbf{V})$, but if moderates protest peacefully alongside them, the situation amounts to a *Revolution* — equilibrium $(\mathbf{P}, \mathbf{V})$. This terminology illustrates the interplay between the size and the intensity of the protests. It is compatible with a variety of political outcomes, which we do not model.

Lemma 1 solves the equilibria of the model under a parametric inequality assumption that guarantees that each of the five types of equilibria exists for some parameter values (see Appendix A.1 for details).

Lemma 1 *There exists a unique equilibrium of the stage game. It is represented on Figure 1 in the (θ_M, θ_R) plane, and fully characterized by the thresholds $\underline{\theta}_R, \overline{\theta}_R$, $\overline{\theta}_M$ and $\underline{\theta}_M$ defined as*

[12] Thus, our solution concept is stronger than Nash equilibrium but weaker than Strong Nash equilibrium (Aumann, 1959), which would also rule out profitable deviations by the grand coalition (all players deviating simultaneously). Strong Nash Equilibria do not exist for all parameter values in our setting.

follows:

$$\begin{cases} \underline{\theta}_R + \alpha \mathbb{E}[\lambda\mu] = \underline{c}, \\ v\overline{\theta}_R + (\gamma - \alpha)\mathbb{E}[\lambda\mu] = \overline{c} - \underline{c}, \\ \underline{\theta}_M + \alpha \mathbb{E}[\mu] = \underline{c}, \\ \overline{\theta}_M + \alpha \mathbb{E}[\mu] - (\alpha + \beta)\mathbb{E}[\lambda\mu] = \underline{c}. \end{cases}$$

The position of θ_R relative to $\overline{\theta}_R$ determines whether radicals prefer collectively playing $a_R = \mathbf{V}$ or $a_R = \mathbf{P}$, whereas the position of θ_R relative to $\underline{\theta}_R$ determines whether radicals are confident enough in their share to start a movement without the moderates. Both parameters are pinned down by radicals' preferences and beliefs about their share.

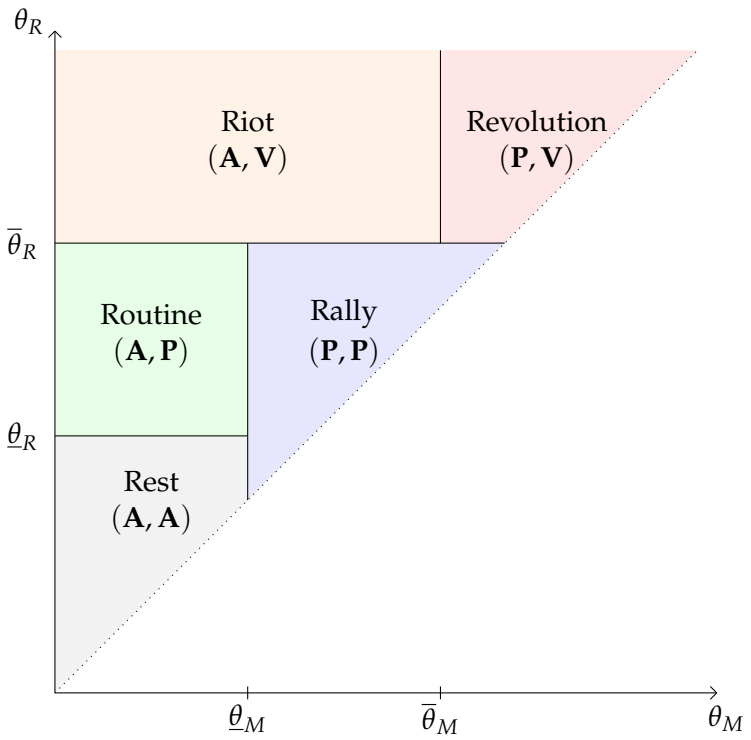
The position of θ_M relative to $\overline{\theta}_M$ (respectively, $\underline{\theta}_M$) determines whether moderates participate or not when radicals protest violently (respectively, peacefully). The strength of preferences required for the moderates to participate when radicals are violent is larger than when radicals are peaceful, as $\overline{\theta}_M \geq \underline{\theta}_M$, illustrating the strategic substitutability at the core of the model — violent movements crowd out peaceful participation.

Comparative statics of the stage game. As a preliminary to the analysis of the role of beliefs in the dynamic model, we perform comparative statics in $\mathbb{E}[(1-\lambda)\mu]$ and $\mathbb{E}[\lambda\mu]$. Consider first an increase in the (perceived) share of moderates, for a fixed number of radicals — that is, an increase in $\mathbb{E}[(1-\lambda)\mu]$ and in $\mathbb{E}[\mu]$ that keeps $\mathbb{E}[\lambda\mu]$ constant. This increase shifts both thresholds $\underline{\theta}_M$ and $\overline{\theta}_M$ downwards. Moderate players become more prone to participation, regardless of the action chosen by radical players. If $\theta < \underline{\theta}_R$, the equilibrium might shift from $(\mathbf{A}, \mathbf{A})$ to $(\mathbf{P}, \mathbf{P})$, thus triggering the participation of radicals as well.

Consider now an increase in the share of radicals, keeping the share of active players constant — that is, an increase in $\mathbb{E}[\lambda\mu]$ and a decrease in $\mathbb{E}[(1-\lambda)\mu]$ for fixed $\mathbb{E}[\mu]$. This increase shifts $\overline{\theta}_R$ and $\underline{\theta}_R$ downwards, and $\overline{\theta}_M$ upwards. Radicals become more prone to participating and resorting to violence, while moderates become less prone to participating if $a_R = \mathbf{V}$. Thus, a regime change from $(\mathbf{P}, \mathbf{P})$ or $(\mathbf{P}, \mathbf{V})$ to $(\mathbf{A}, \mathbf{V})$ is possible: the fact that radicals resort to violence and are perceived to be more numerous crowds out moderates' participation. Lemma 2 summarizes these results: Large protests arise if the perceived number of moderates is large, and if the perceived number of radicals is small.

Lemma 2 *Fix all parameters of the game except for beliefs. Then:*

Figure 1: Equilibria of the stage game: the 5R of revolts



Note: Recall that $\theta_R > \theta_M$. This picture assumes that $\underline{\theta}_R < \overline{\theta}_R$, as explained in the appendix. If $\underline{\theta}_R \geq \overline{\theta}_R$, the green region disappears, and $(\mathbf{A}, \mathbf{P})$ is not a possible equilibrium, as radicals always prefer coordinating on violent action than protesting peacefully. The picture also assumes that $\overline{\theta}_R > \overline{\theta}_M$, but the opposite inequality is also possible (and all five regions still exist in that case).

(i) Fixing $\mathbb{E}[\lambda\mu]$, there exists a threshold m such that a large protest arises if and only if $\mathbb{E}[\mu] \geq m$.

(ii) Fixing $\mathbb{E}[\mu]$, there exists a threshold r such that a large protest arises if and only if $\mathbb{E}[\lambda\mu] \leq r$.

2.2 Dynamics of protests

Beliefs about the population's preferences influence individuals' decisions to protest. Conversely, protest movements reveal information about the population's preferences. In this section, we analyze the joint evolution of beliefs and political participation in a dynamic framework.

2.2.1 Dynamic framework

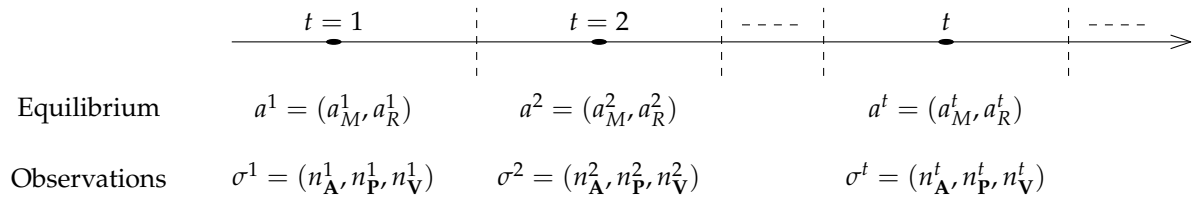
Timeline. The stage game described in subsection 2.1 is played at each period of an infinite horizon. Time is discrete and indexed by $t \in \{1, 2, \cdots\}$. Players are short-lived or, equivalently, myopic.

All players start the game with a common prior belief over (λ, μ) described by the full-support pdf $\chi_0 : [0,1]^2 \to [0,1]$.[13] We write (λ, μ) for the generic variable and $(\tilde{\lambda}, \tilde{\mu})$ for the correct value. Since agents are short-lived, at each date t, they play an equilibrium of the stage game given their beliefs $\chi_t(h_t)$, where χ_t is the Bayesian posterior following history h_t. We write $a^*(\chi) = [a_M^*(\chi), a_R^*(\chi)]$ for the equilibrium of the stage game under belief χ.

Information. After each date t, the behavior of n players at the last stage game is publicly displayed. These n players are randomly, uniformly and independently selected from the population. That is, the probabilities with which a selected individual is passive, moderate or radical equal $1 - \tilde{\mu}$, $(1 - \tilde{\lambda})\tilde{\mu}$ and $\tilde{\lambda}\tilde{\mu}$, respectively.

A history h_t at date t therefore consists, for each date s up to t, of: (i) the nature of the stage-game equilibrium played at s, represented by $a^s = (a_M^s, a_R^s) \in \{\mathbf{A}, \mathbf{P}\} \times \{\mathbf{A}, \mathbf{P}, \mathbf{V}\}$; (ii) the number n_a^s of individuals playing action $a \in \{\mathbf{A}, \mathbf{P}, \mathbf{V}\}$ at date s, where $n_{\mathbf{A}}^s + n_{\mathbf{P}}^s + n_{\mathbf{V}}^s = n$. We write $\sigma = (n_{\mathbf{A}}, n_{\mathbf{P}}, n_{\mathbf{V}})$ generically for the signal, and $f(\sigma | a, \tilde{\lambda}, \tilde{\mu})$ for the actual signal distribution conditional on the equilibrium a being played and on the true preference parameters being $(\tilde{\lambda}, \tilde{\mu})$. We also abuse notation and write $\chi(\sigma \mid a)$ for the belief over the signal that is implied by the equilibrium a and the distribution χ over (λ, μ), and $\mathbb{E}_\chi[y]$ for the subjective expected value of variable y under belief χ.

Figure 2: Timeline



[13] The fact that χ_0 has full support implies that agents' models are correctly specified, and hence learning the correct values of λ and μ is theoretically possible. This distinguishes our model from the literature on misspecified learning (e.g. Esponda and Pouzo, 2016; Bohren and Hauser, 2021), where convergence is impeded by a prior that assigns null weight to the true value.

Equilibrium concept. We analyze the long-run outcomes that result from the co-evolution of beliefs and actions, with a particular interest in situations where learning about the population's preferences is incomplete. To do so, we compare two objects: (i) the *full-information equilibrium*, that is, the equilibrium that would be played if all players knew $\tilde{\lambda}$ and $\tilde{\mu}$; (ii) the possible *long-term equilibria* achieved once actions and beliefs have converged. We model the latter as the set of *self-confirming* equilibria (Fudenberg and Levine, 1993). Formally:

Definition 1 *A self-confirming equilibrium is a triple* $[a, \chi, (\tilde{\lambda}, \tilde{\mu})]$ *such that* $(\tilde{\lambda}, \tilde{\mu}) \in supp(\chi)$ *and:*

$$\begin{cases} a = a^*(\chi), \\ \chi(. \mid a) = f(. \mid a, \tilde{\lambda}, \tilde{\mu}). \end{cases}$$

A self-confirming equilibrium restricts beliefs and actions to be consistent with each other on the path. The first condition states that the population plays the equilibrium prescribed by the belief χ. The second condition states that beliefs are ultimately correct on the equilibrium path: the rationale is that, if a is played infinitely often, beliefs about the frequency of equilibrium actions should converge to the correct value, as individuals have access to an infinite sample from the population playing a. Importantly, the population might maintain incorrect beliefs about off-path events. Standard results from the literature on active learning (Fudenberg and Levine, 1993) imply that: (i) when playing the repeated game, society almost surely converges on an equilibrium, which must be a self-confirming equilibrium; (ii) conversely, any self-confirming equilibrium can be reached asymptotically with positive probability from an appropriate prior.

Learning traps. Our main interest lies in situations where information about the population's preferences is imperfectly revealed asymptotically, yielding an equilibrium that differs from the full-information equilibrium. We call these situations *learning traps*. Let $\delta_{\tilde{\lambda},\tilde{\mu}}$ be the Dirac distribution on $(\tilde{\lambda}, \tilde{\mu})$.

Definition 2 *A learning trap is a self-confirming equilibrium* $[a, \chi, (\tilde{\lambda}, \tilde{\mu})]$ *such that* $a \neq a^*(\delta_{\tilde{\lambda},\tilde{\mu}})$.

In a learning trap, individuals end up forming correct beliefs about their payoffs in the long-run equilibrium they play, but they misperceive the share of radicals or moderates in the population. As a result, they keep incorrect beliefs about the payoffs they would receive if different actions were played.

Learning traps arise when the equilibrium profile of actions played in the long run does not reveal enough information for players to revise incorrect beliefs. The simplest example is the one where a protest is never initiated — $(\mathbf{A}, \mathbf{A})$ is played forever — out of undue pessimism about the population's preferences. The same logic applies in other situations where one group of players underestimate their share and thus fail to activate their margin of participation (e.g., $(\mathbf{A}, \mathbf{P})$ instead of $(\mathbf{P}, \mathbf{P})$ due to moderates' pessimism about μ, or $(\mathbf{P}, \mathbf{P})$ instead of $(\mathbf{P}, \mathbf{V})$ due to radicals' pessimism about λ). By definition, the profile $(\mathbf{P}, \mathbf{V})$ cannot be played in a learning trap, as it would reveal the values of $\tilde{\lambda}$ and $\tilde{\mu}$, yielding rational expectations. Conversely, if the full-information equilibrium is $(\mathbf{A}, \mathbf{A})$, then it is reached with probability one from any correctly specified prior.

As standard in the literature on protests, learning traps are therefore asymmetric: information frictions can systematically hinder the coordination that is necessary to give rise to large-scale protests. Participation is therefore typically lower on at least one margin (intensive or extensive) in a self-confirming equilibrium than in the corresponding full-information equilibrium.

However, our framework offers one exception to this logic due to the strategic substitutability between both types of protesters: it is possible that the game converges on $(\mathbf{P}, \mathbf{P})$ while radicals, who underestimate their share, would trigger a violent action and crowd out moderates, yielding equilibrium $(\mathbf{A}, \mathbf{V})$, if they had correct beliefs. In that case, information frictions modify the nature of the social movement by affecting the intensive and extensive margins in opposite directions: radicals' failure to coordinate is the only thing that prevents a lasting peaceful movement from turning into a riot. Table A.1 in Appendix A.3 gives a complete characterization of all possible learning traps, and Proposition 1 summarizes them.

Proposition 1 *There are two categories of learning traps:*

(i) those that reduce the extensive or intensive margin of protests (or both) due to an underestimation of $\tilde{\lambda}\tilde{\mu}$, of $\tilde{\mu}$ (or of both);

(ii) a learning trap where the game converges on a Rally $(\boldsymbol{P}, \boldsymbol{P})$ while the full-information equilibrium would be a Riot $(\boldsymbol{A}, \boldsymbol{V})$.

2.2.2 Participation costs and the persistence of protests

We now turn to exploring how variations in $\underline{c}$ and $\overline{c}$ affect the persistence of protests. This question is important for two applications. First, repression by the government can

be seen as an attempt to increase the cost of protesting in order to discourage participation. Second, social media, which we study more specifically in Section 2.3, can be thought of as an alternative realm where expression of discontent is possible at a lower personal cost than in the streets.

If participation decisions are strategic complements, reducing the cost of participation always encourages more players to join the movement. If a large protest reveals strong popular support, this can further trigger a sustainable movement by allowing the players to escape a learning trap with insufficient activity.

In our framework, moderates' participation thresholds $\underline{\theta}_M$ and $\overline{\theta}_M$ (on Figure 1) are increasing in $\underline{c}$, as in existing models. But radicals' equilibrium level of violence depends on the relative cost $\overline{c} - \underline{c}$, and moderates' participation is crowded out by radicals' violence. A reduction in the costs of participation, e.g. due to social media, that also decreases the relative cost of violence $\overline{c} - \underline{c}$, therefore makes radicals more prone to playing $a_R = \mathbf{V}$. If $\theta_M \in (\underline{\theta}_M, \overline{\theta}_M)$, this variation can also affect the extensive margin of protests, switching from a *Rally* $(\mathbf{P}, \mathbf{P})$ to a *Riot* $(\mathbf{A}, \mathbf{V})$, as moderates now abstain from participating alongside violent protesters. Lower participation costs are thus a double-edged sword because they make it easier for all factions to coordinate, including radicals.

In other words, a reduction in the cost of radical action might eliminate the learning trap in which incomplete information is the only thing that precludes the rise of a violent movement (item (ii) in Proposition 1). Social media can trigger a large protest, but it can also cause its demise: the former by helping moderates to coordinate, the latter by helping radicals to coordinate. Conversely, a government trying to nip a protest movement in the bud by intensifying police repression can paradoxically favor a large protest by preventing its radicalization. The key statistic that controls this comparative statics is whether variations in costs affect $\overline{c}$, the cost from violent action, proportionately more or less than $\underline{c}$, the cost from peaceful protesting — for instance, whether repression focuses mostly on violent protesters or cracks down indiscriminately on all participants.

In Proposition 2 we formalize these observations by studying the space of parameters (beliefs and true shares) conducive to a self-confirming equilibrium with a large protest. The proposition reveals that, in contrast with a game of strategic complements only, a decrease in the costs of participation does not necessarily make large protests easier to sustain.

Proposition 2 *Let $\Omega(\underline{c}, \overline{c})$ be the set of all pairs $[\chi, (\tilde{\lambda}, \tilde{\mu})]$ such that $[a, \chi, (\tilde{\lambda}, \tilde{\mu})]$ is a self-confirming equilibrium under participation costs $(\underline{c}, \overline{c})$ for some large protest $a \in \{(\boldsymbol{P}, \boldsymbol{P}), (\boldsymbol{P}, \boldsymbol{V})\}$. Let $\underline{c}' < \underline{c}$ and $\overline{c}' < \overline{c}$. Then:*

(i) if $\overline{c}' - \underline{c}' \geq \overline{c} - \underline{c}$*, then* $\Omega(\underline{c}', \overline{c}') \supset \Omega(\underline{c}, \overline{c})$ *for all* $(v, \theta_M, \theta_R, \alpha, \beta, \gamma)$*;*

(ii) if $\overline{c}' - \underline{c}' < \overline{c} - \underline{c}$*, then there exist values of* $(v, \theta_M, \theta_R, \alpha, \beta, \gamma)$ *such that* $\Omega(\underline{c}', \overline{c}') \not\supseteq \Omega(\underline{c}, \overline{c})$.

2.3 Social media: a catalyst for protest dynamics

According to our framework, lower participation costs have ambiguous effects on protest size. In this section, we use social media as an exogenous shift in participation costs that reveals information about the population's preferences at lower cost than street protests do. This assumption allows us to make sharper predictions regarding the impact of a low-cost mobilization technology on protest dynamics in the short run. Then, we discuss the robustness of our findings if we consider other characteristics of social media: their vulnerability to government control, and their propensity to deliver biased information.

To integrate social media in the model, we modify the timeline in Figure 2 by dividing each period t into two subperiods: at ta, individuals make *online* participation decisions; at tb, they make *offline* participation decisions. After each subperiod ta or tb, the number of players selecting each possible action $\{\mathbf{A}, \mathbf{P}, \mathbf{V}\}$ among n randomly selected individuals is revealed to all subsequent cohorts.

The payoffs to online participation decisions are given by Equations (1) and (2), except that participation costs are equal to $\underline{c}' < \underline{c}$ and $\overline{c}' < \overline{c}$, where $\underline{c}$ and $\overline{c}$ still measure offline cost parameters. We assume that $\overline{c}' - \underline{c}' < \overline{c} - \underline{c}$: the relative cost of violent participation is smaller on social media than in the streets. The long-run effects of social media are thus captured by the comparative statics in $\underline{c}$ and $\overline{c}$ performed in the previous section. In this section, we instead analyze finite-time dynamics to understand how social media affects the birth, momentum and decline of a protest movement.

2.3.1 Crowd-in-then-crowd-out sequence

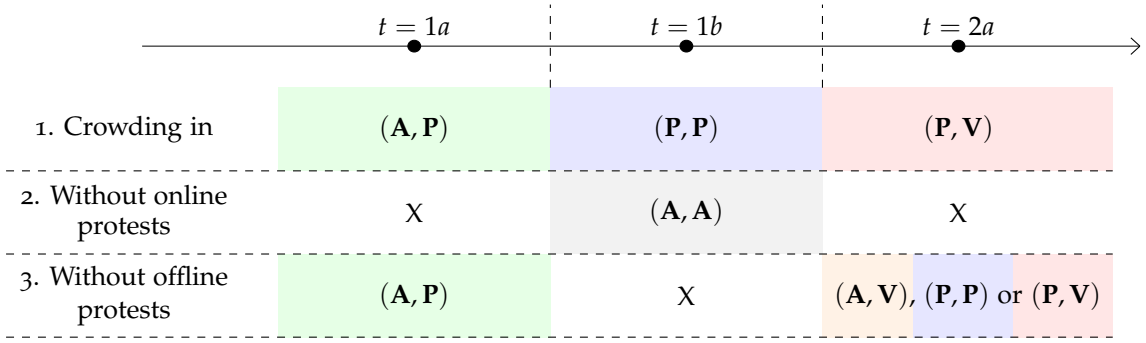
A large number of sequences are possible for any prior because signals are stochastic. Therefore, instead of providing a comprehensive categorization of all possible sequences, we use the model to elucidate a specific dynamics where social media plays a crucial role in both the initial, peaceful emergence of a social movement, and in its radicalization and demise. We decompose this sequence into two stages: an initial *crowding in*, where online participation increases offline participation and vice versa, and a subsequent *crowding out*, where radical activity causes moderates to leave the movement. In Section 3 we

provide evidence for all stages of this dynamics in the case of the Yellow Vest movement. We now explain how to interpret this pattern in light of our conceptual framework.

To capture the situation where social media is instrumental in the launch and initial momentum of a protest movement, we assume that, in the absence of social media, neither margin of participation is ever activated.

Crowding in. Consider first the crowd-in phase, illustrated in Figure 3 (sequence 1). Due to the lower cost of online mobilization, social media initiates an online movement (equilibrium $(\mathbf{A}, \mathbf{P})$ in period $1a$) where participation is larger than expected. This makes players more optimistic about the population's preferences, triggering a massive but peaceful offline protest in period $1b$ (equilibrium $(\mathbf{P}, \mathbf{P})$). Consider now a counterfactual scenario where, *ceteris paribus*, the cost of protesting online in $1a$ is large enough to discourage mobilization (sequence 2). This prevents learning in period $1a$, and no protest takes place in $1b$. We document this causal effect of online activity on offline protests in Section 3.2, by showing that the presence of a 4G antenna near a municipality increases participation online and in subsequent offline protests.

Figure 3: Crowding-in and counterfactual scenarios

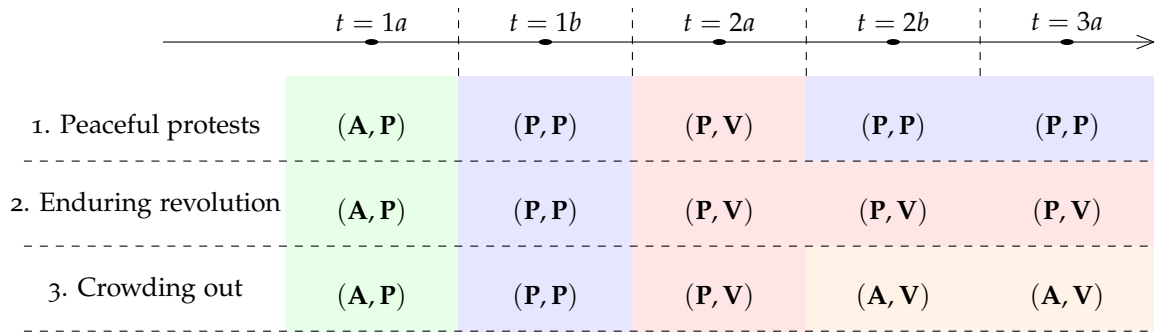


The massive protest $(\mathbf{P}, \mathbf{P})$ in period $1b$ then reveals a large value of μ, which encourages larger and more radical mobilization online in period $2a$ (equilibrium $(\mathbf{P}, \mathbf{V})$). In a counterfactual world without a possibility of offline activity in $1b$ (sequence 3), the disclosure of information about μ is eliminated, and hence the online equilibrium in period $2a$ is determined only based on the information revealed online in period $1a$. The absence of information revelation at $t = 1b$ can then delay or prevent the advent of a large online protest (if $(\mathbf{A}, \mathbf{V})$ is played in $t = 2a$). In Section 3.2 we provide evidence for the causal effect of offline mobilization on the size of online mobilization by showing that municipalities located near a highway toll (which served as coordination devices for Yellow Vest protesters and reduced the cost of offline activity) experienced higher online

activity in the subsequent period. We complement this finding by showing descriptively that online mobilization was more radical on Facebook pages created after the 11/17 protests.

Crowding out. The Yellow Vest movement then exhibited a decline in participation and an increase in violent actions. We ask what this dynamics reveals by comparing it to two alternative evolutions of the movement following the initial crowding in. Our key insight is that the parameter that governs the evolution of the movement is the information revealed about the share of radicals. To illustrate this, consider the three sequences in Figure 4.[14] In these sequences, the movement has first gained momentum, as $(\mathbf{A}, \mathbf{P})$, $(\mathbf{P}, \mathbf{P})$ and $(\mathbf{P}, \mathbf{V})$ are played from $t = 1a$ to $t = 2a$. This happens when the population learns from online interactions that the propensity to mobilize μ is larger than expected, encouraging a massive protest in period $1b$. Besides, radicals become more optimistic about their share, prompting them to resort to radical online expression in period $2a$.

Figure 4: Diverging sequences after initial crowding-in.



In sequence 1, street protests never turn violent. An illustration is provided by the non-violent 2014 Umbrella Movement in Hong Kong, which lasted several months (see, e.g., Cantoni et al., 2019). Sequence 2 corresponds to a case where social media helps organize massive protests that turn into enduring revolutions. This sequence is compatible with the Arab Spring in the early 2010s, which began as a local protest in Tunisia and led to massive unrest ranging from demonstrations to civil war in more than fifteen countries (see, e.g., Steinert-Threlkeld, 2017; Acemoglu et al., 2018; Brummitt, Barnett and D'Souza, 2015). Last, sequence 3 shows the *crowding-out* dynamics where, as in the

[14] Proposition 3 predicts the equilibrium in period $t = 2b$ only. For simplicity, on Figure 4 we assume that the equilibrium in period $t = 3a$ is the same. Our analysis of the Yellow Vest movement in Section 3.3 indeed reveals some online radicalization after initial crowding out.

case of the Yellow Vests, moderates leave the movement. In Section 3.3, we provide evidence that radical expression on online Yellow Vests discussion pages indeed decreased subsequent participation by moderate individuals.

We elucidate the conditions under which either of the three sequences depicted in Figure 4 obtains. Consistently with Lemma 2, Proposition 3 confirms that the fate of the movement following initial crowd-in depends on what online activity reveals about the share of radical participants.

Proposition 3 *Suppose that the game starts with a crowding-in phase where* $(\boldsymbol{A},\boldsymbol{P})$, $(\boldsymbol{P},\boldsymbol{P})$ *and* $(\boldsymbol{P},\boldsymbol{V})$ *are played at* $t = 1a$, $t = 1b$ *and* $t = 2a$, *and that* $\theta_M + \alpha\mathbb{E}_{\chi_{2b}}[\mu] \geq \underline{c}$. *Then, there exists thresholds* $r_1 \leq r_2$ *such that the subsequent equilibrium profile of actions at* $t = 2b$ *is determined as follows:*[15]

(i) if $\mathbb{E}_{\chi_{2b}}[\lambda\mu] < r_1$, *then* $(\boldsymbol{P},\boldsymbol{P})$ *is played (peaceful protest);*

(ii) if $r_1 < \mathbb{E}_{\chi_{2b}}[\lambda\mu] < r_2$, *then* $(\boldsymbol{P},\boldsymbol{V})$ *is played (enduring revolution);*

(iii) if $r_2 < \mathbb{E}_{\chi_{2b}}[\lambda\mu]$, *then* $(\boldsymbol{A},\boldsymbol{V})$ *is played (crowding out).*

2.3.2 Extensions and discussions

Digital repression. Governments use a variety of instruments to control protests. An important example is the shutting down of social media. This instrument is both used by authoritarian regimes to restrain legitimate democratic movements, and by democratic regimes to contain violent protests.[16]

Our model offers a framework for considering the potential consequences of a social media shutdown. The effects of this policy depend on when it is implemented. Consider the crowd-in-then-crowd-out sequence to see this. Implementing a ban on social media from period $t = 1a$ onward would prevent the movement from gaining momentum. This would result in the lowest participation equilibrium, $(\mathbf{A},\mathbf{A})$, in all future periods. However, implementing the ban from period $t = 2a$ onward could prevent the radicalization of the movement and subsequent crowding out of moderates. This would result in equilibrium $(\mathbf{P},\mathbf{P})$ in all future periods.

[15] The interval (r_1, r_2) might be empty, in which case $(\mathbf{P},\mathbf{V})$ is never played.

[16] The Iranian regime imposed nationwide internet blackouts during the November 2019 and January 2026 protests. According to the Centre for International Policy Studies, nearly half of the Internet shutdowns in Africa in 2022 were imposed during political unrest. In 2024, the French government blocked TikTok in the overseas territory of Nouvelle-Calédonie, which was the scene of violent riots. On the other hand, we are not aware of any similar actions taken during the Yellow Vest movement.

Paradoxically, shutting down social media once the movement has gained traction then favors its persistence, even in the absence of a specific reaction by protesters against the shutdown. This mechanism fits well with the observation that many shutdowns are actually followed by an escalation of the momentum of preexisting protests, or at least a continuation of past dynamics (see, for example, Rydzak, Karanja and Opiyo (2020) in the case of protests in several African countries between 2017 and 2019). The strategic analysis of the optimal policy for a government that aims to contain peaceful and/or violent protests must therefore consider the effects on both margins.

Second, authoritarian governments can control speech on social media by cracking down on online anonymity and repressing the expression of political opinions. In such contexts, the assumption that $\underline{c}' < \underline{c}$ and $\overline{c}' < \overline{c}$ is unrealistic. If it is more costly to express both radical and moderate expressions online than offline, then social media loses its coordinating power. If, however, only radical expression is repressed ($\overline{c}' > \overline{c}$ and $\underline{c}' < \underline{c}$), online repression can backfire and favor the persistence of a larger protest by hindering the radicalization and crowding out that would happen otherwise.[17]

Algorithmic bias. Our analysis so far assumes that the information received by the population is unbiased, in that it accurately reflects the shares of the different types of protesters. However, the algorithms used by social media platforms may skew the content shown to users. This is consistent with existing experimental evidence (see Levy, 2021) and our own descriptive analysis of Yellow Vest Facebook discussion pages (see Section 3.3).

In our model, an over-sampling of radical content on social media leads individuals to overestimate $\lambda\mu$. To understand the consequence of this overestimation, consider a movement that has gained momentum, and where $(\mathbf{P}, \mathbf{V})$ is played online for the first time, as in period $t = 2a$ of Figure 4. Following the comparative statics of Proposition 3, a bias towards radical content at $t = 2a$ encourages radical actions in $t = 2b$, possibly at the cost of smothering a large protest. This phenomenon, which is a byproduct of the strategic substitutability between moderates and radicals, qualifies the common wisdom that sees social media bias as a necessary catalyst of massive protest movements. In the case of the Yellow Vest uprising, to which we now turn, this bias may have accelerated the demise of the movement.

[17] Similar arguments could be used to analyze censorship or self-policing as in Shadmehr and Bernhardt (2015) and Ananyev, Xefteris, Zudenkova and Petrova (2019).

3 Empirical application: the Yellow Vest movement

In this section, we analyze the Yellow Vest movement through the lens of our theoretical framework. More specifically, we present several pieces of evidence consistent with the crowd-in-then-crowd-out sequence studied in Section 2.3.

3.1 Context and data

While the Yellow Vest movement is linked to longstanding and growing discontent over spatial inequalities and related environmental policies (Algan, Beasley, Cohen, Foucault and Péron, 2019; Boyer, Delemotte, Gauthier, Rollet and Schmutz, 2020; Douenne and Fabre, 2022), its timing and widespread initial success were largely unexpected. It was sparked by an online petition and quickly organized on social media. The first week of protests took the form of hundreds of roadblocks across France. Then, for a few months, more traditional protests took place every week in medium and large cities. However, protests quickly turned violent (in the days following the first protests), drew fewer participants, and eventually disappeared. We provide more elements of context in Appendix B and additional information on our data in Appendix C.

3.1.1 Data

Sources. To understand the roots of the movement, we obtained anonymized geolocated data from Change.org on the timing of petition signatories through the end of 2019. To proxy for offline mobilization, we collected a map of planned roadblocks on the evening of November 16th, 2018. The map was downloaded directly from a website created by protesters to coordinate demonstrations and roadblocks. It documented 788 announced roadblocks in metropolitan France, all of which pointed to precise road infrastructure (e.g., highway access ramps and tolls, parking lots, roundabouts) and included specific descriptions of the planned events. Given the nature of this source, our measure provides a lower bound of the extent of mobilization, focused on the largest protests.[18] Many locations were chosen for their potential to block traffic and economic

[18]Two potential concerns arise with our data. First, protesters might have falsely declared intent to demonstrate. This is unlikely: since the map was created to coordinate roadblocks, there was little incentive to overstate participation. Moreover, unlike in autocratic regimes (Clarke and Kocak, 2020; Hassan, 2021), the French police did not preemptively try to lift the roadblocks. We gathered actual press reports (from the universe of daily newspapers in the two days following 11/17) on 613 of these announced roadblocks, and we also use this more conservative measure for robustness (see Appendix

activity. Based on the division of the country into *Bassins de vie* (hereafter referred to as Living Zones), we estimate that more than half of the country's population and more than a third of the country's territory were directly affected by a roadblock.[19] We complement this data with official weekly statistics on the number of protesters.[20]

To document the online equivalent of street protests, we searched for all public Facebook groups related to the movement. Using the methodology of Gaby and Caren (2012), we compiled a list of the Facebook groups that were still active one month after 11/17 by performing search requests using a large set of keywords linked to the movement. We recorded each group's name, creation date, number of members, and publications. We identified 3,033 groups with a total of four million members. Over two-thirds of the groups were associated with a geographical area, and more than 40% of the total members belonged to these localized groups. Moreover, only 20% of the posts emanated from national groups, suggesting that localized groups were the most active ones. Using a similar method, we also identified 617 public Facebook pages and used Netvizz (Rieder, 2013) to retrieve their content in March 2019. This corpus features 120,227 posts, 2.1 million comments, 2.8 million sentences, and 21 million interactions (likes and reactions). Since Netvizz did not provide discussant identifiers associated with each message, we scraped Facebook a second time in January 2022 and enriched the dataset with 120,463 distinct discussant identifiers for 377,283 messages and 706,165 sentences.[21]

Table D.1). Second, our source might undercount protest locations by missing smaller, spontaneous events. The Ministry of the Interior announced 2,034 protest sites during the first day of protests, but these statistics are not made available to researchers. However, this figure is consistent with our data. The 788 points on the map are usually at the municipal level, and 300 of them include short descriptions of the protest. Among those 300, each lists an average of 2.44 distinct blockade sites within the municipality. Assuming this pattern holds for all points on the map, a rough back-of-the-envelope calculation on our data implies 1,922 distinct blockade sites. In practice, we aggregate the data at the municipal level, so any undercounting of blockade sites within a municipality does not affect our analysis.

[19] Living Zones are statistical units defined as the smallest groups of municipalities where residents have access to basic services and can conduct a large part of their daily lives. According to our data, 551 of the 1,632 Living Zones were affected.

[20] Protests took place on Saturdays. Estimates of the 11/17 protests range from 287,700 (Ministry of the Interior) to 1.3 million (a police union). We choose to report the official statistics to ensure consistency of the time series.

[21] To protect users' privacy, all users were de-identified. Approximately 30% of pages had been deleted by January 2022 (see Appendix Table C.2). To assess selection bias, we extensively compared both datasets. They are similar in terms of their distribution of political language and in terms of the topics discussed (see Appendix Figure E.5).

Textual analysis of Facebook Discussions. To analyze discussions on Facebook pages, we rely on text-as-data methods (see, for an overview, Grimmer and Stewart (2013), Gentzkow et al. (2019) and Ash and Hansen (2023), and Appendix E for details). Our preferred method is a topic model tailored to analyze short text snippets (Demszky, Garg, Voigt, Zou, Gentzkow, Shapiro and Jurafsky, 2019). Among our topics, some relate to protest organization, socialization, and online mobilization. Others reflect the reasons behind the protests and the political goals the Yellow Vests were trying to achieve. Finally, several topics refer to antagonistic messages and reflect the protesters' anger toward government officials and their policies. In what follows, we will focus on results associated with the probability that any given sentence is associated with an antagonistic topic, and use other measures of radicalism for robustness.[22]

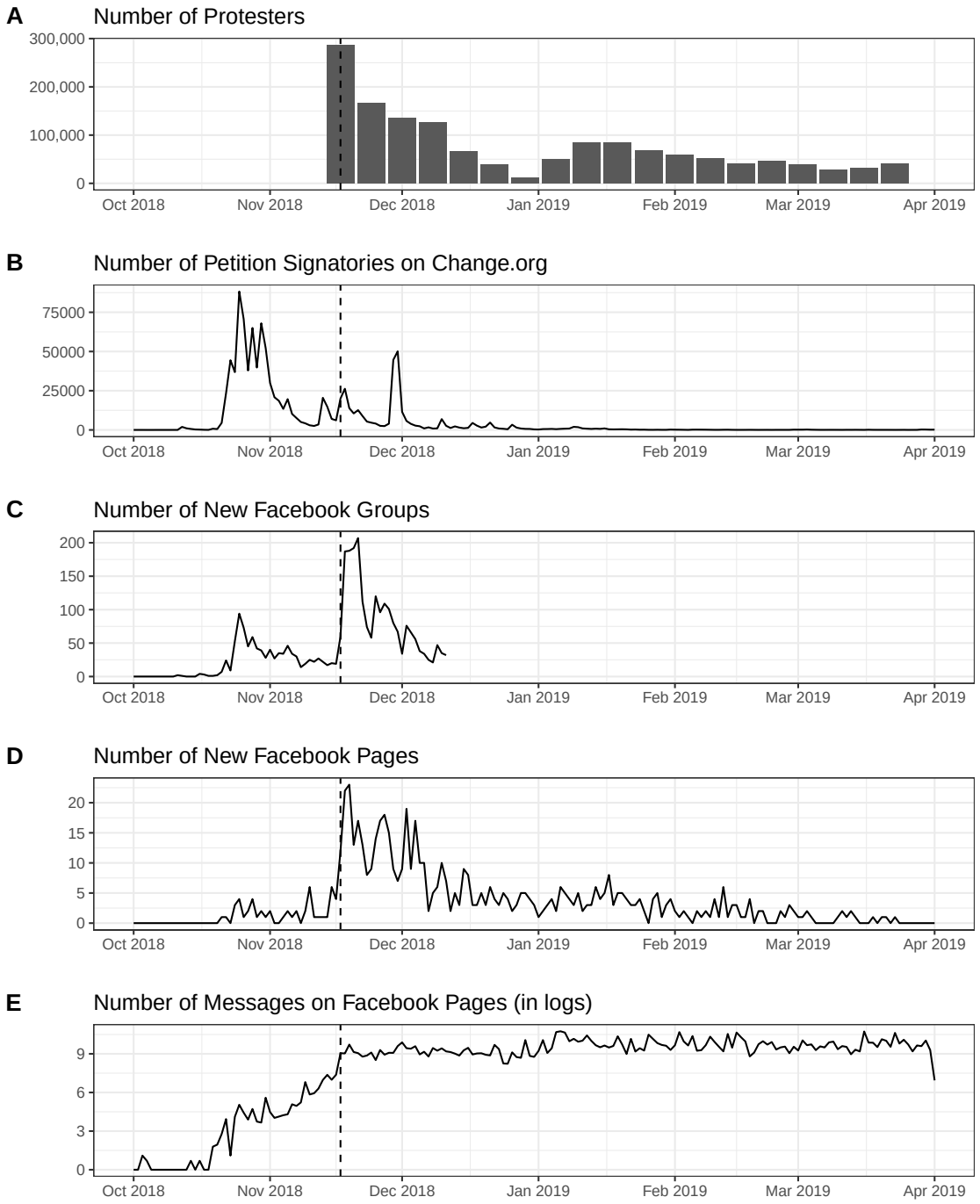
3.1.2 Time series

In Figure 5, we combine the weekly time series of the official number of Yellow Vest protesters on the streets with the daily time series of the number of petition signatures, the number of Facebook group and pages creations, and the number of comments on Facebook pages. The movement culminated in the streets during the first episode of the protests (Panel A). While most of the signatures on the petition were collected before 11/17 (Panel B), there were two distinct episodes of group creation: one in the weeks before 11/17 and one immediately after (Panel C). This illustrates social media's dual role as a coordination/counting device and a means of expression: Facebook groups were used to organize the roadblocks, but also served as virtual meeting places that allowed the movement to continue after the initial street mobilization. The creation of many discussion pages after 11/17 supports this hypothesis (Panel D). These pages remained very active in the following months, in sharp contrast with the decline in the weekly number of protesters in the street (Panel E).

Protest violence. To measure the evolution of street violence, we use the number of crimes related to rioting (arson, destruction, and fighting with officers) recorded by the police from 2000 to 2019. In Panel A of Figure 6, we plot the residual of a regression

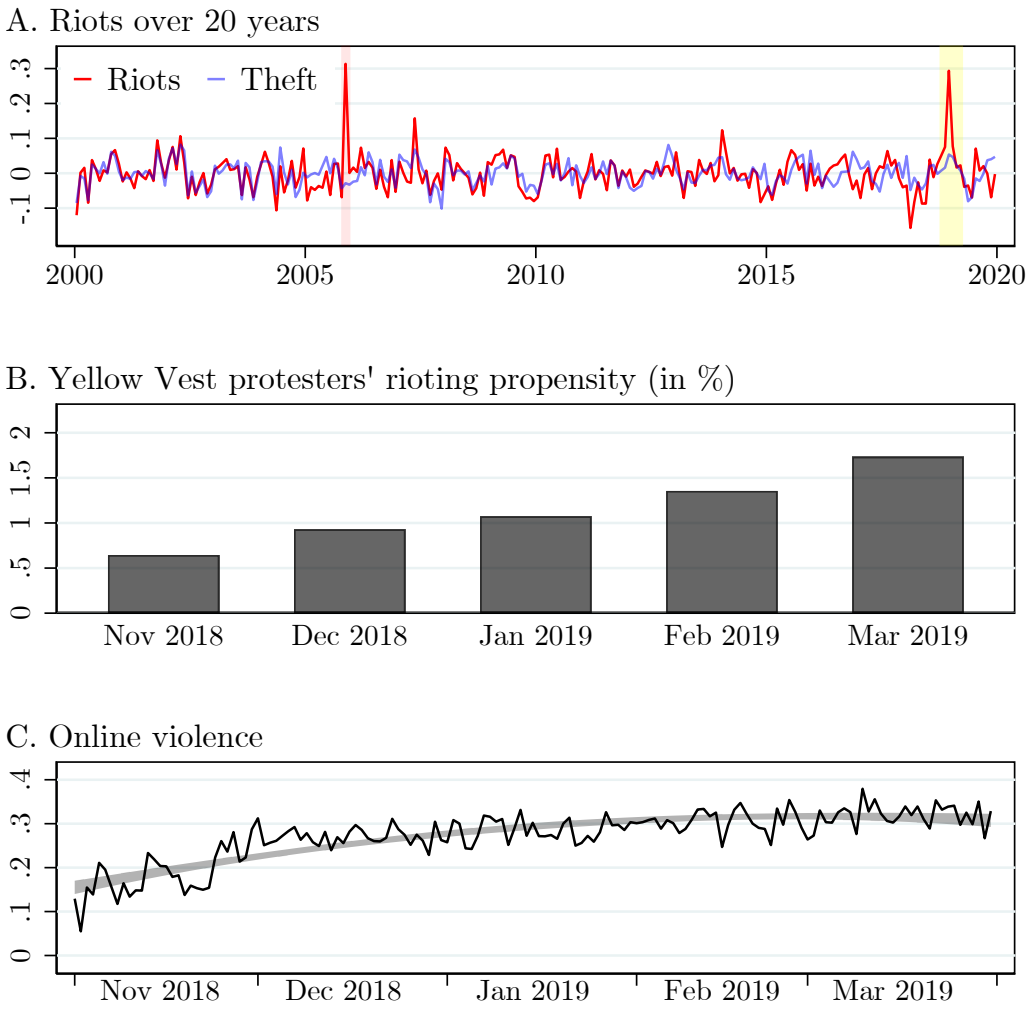
[22]To measure the emotional content of messages, we use a dictionary-based approach that assigns a sentiment score to each sentence and focus on negative sentiment. To understand messages' political stance, we train a supervised learning model that predicts the party affiliation of members of the French Parliament based on their tweets and use it to construct the probability of a given sentence being written by affiliates of either far-left or far-right parties.

Figure 5: Evolution of Online and Offline Mobilizations



Notes: Panel A reports the weekly number of demonstrators as recorded by the Ministry of the Interior. Panel B shows the daily number of petition signatures. Panels C and D show the daily number of new Facebook groups and Facebook pages created, respectively. Panel E shows the daily number of messages posted on Facebook pages (in logs). The vertical dashed line in each panel corresponds to 11/17.

Figure 6: Evolution of Online and Offline Violence



Notes: In Panel A, we show in red the residual from an OLS regression of the monthly log number of riot-related offenses reported by the Ministry of the Interior, after controlling for month fixed effects. For comparison, we replicate this analysis for theft-related offenses, which are three times more numerous (in blue). Vertical bars highlight the November 2005 riots and the Yellow Vest protests (November 2018-March 2019). In Panel B, we divide the hypothetical number of riot-related offenses that may be plausibly attributed to the Yellow Vests (10% of riot-related offenses, which corresponds to the average value for the period highlighted in yellow in Panel A) by the number reported in Panel A of Figure 5. In Panel C, we show the daily share of sentences in our text corpus that belong to an antagonistic topic, with a quadratic fit.

of the monthly log number of riot-related offenses on month fixed effects, to control for seasonality. Street violence increased sharply between November and December 2018, despite a steady decline in the number of protesters. In December, it was comparable to that of the November 2005 youth riots, which were the most intense riots in France since 1968. To get a sense of the average rioting propensity of Yellow Vest protesters, we

hypothesize that 10% of riot-related offenses over our study period are linked with the movement.[23] We divide this index of over-rioting by the number of protesters displayed in Panel A of Figure 5 and plot the resulting ratio in Panel B, which suggests that the average rioting propensity of a Yellow Vest protester tripled from November 2018 to March 2019.

Finally, we show in Panel C that online violence followed a similar trend. To measure the evolution of online violence, we conduct our textual analysis between the end of October 2018 and the beginning of April 2019. Our topic model shows that the share of messages associated with political or economic concerns decreased, while messages of violence, conspiracy theories, and insults increased (see Appendix Figure E.2). Overall, the share of messages associated with antagonistic content doubled from 15% to 30% over the period.[24] For lack of a better word and to stay close to the terminology used in Section 2, we refer to the rise of online violence under the umbrella term *radicalization*.

3.2 Crowding in: the online-offline feedback loop

We first present evidence consistent with the crowd-in sequence depicted in Figure 3 for periods 1a, 1b and 2a, namely: (i) Early online mobilization helped organize the first street protests; (ii) These first protests led to an increase in the size of online activity. We first assess the effect of early online mobilization (before 11/17) on the occurrence of a protest on 11/17. According our framework, finding a positive effect would mean that protesters were initially too pessimistic about the potential of the mobilization to start a street protest from scratch, but were able to revise their priors upwards by first mobilizing online. Then, we turn to the reverse direction and assess the effect of the 11/17 protests on the extent of subsequent online mobilization. While there are many

[23] While street violence decreases after December, this pattern was partly driven by heavy police response, on which we have no reliable information (see Petrovskii, Shishlenin and Glukhov (2025) for a discussion). Street violence was, on average, 10% higher than the 2000-2019 mean between November 2018 and March 2019.

[24] Similarly, the share of messages classified as negative sentiment or politically-extreme also increased, albeit to a lesser extent (see Appendix Figure E.7). While negative sentiment could encompass very different emotions, we provide suggestive evidence that anger drove this increasing pattern (see Appendix Figure E.4). Of course, some messages that contain antagonistic elements or show negative sentiments may also reflect the fact that online discussants are describing violent events that they witnessed or were victims of in the streets, without necessarily endorsing violence themselves. However, our third classification based on partisan affiliation is less subject to this potential bias, consistently with polling data showing that the decline in popular support for the movement was mostly driven by centrist voters (see Appendix Figure C.1).

possible explanations for this pattern,[25] our framework suggests that higher subsequent online mobilization is indeed more likely after a large street protest has revealed that a higher share of the population was prone to mobilizing.

3.2.1 Empirical strategy

In the absence of individual-level information on both online and offline mobilization, we construct a dataset at the most granular level possible: the municipality. Municipalities represent the lowest tier of government and a wide range of social, economic, geographical and political characteristics, listed in Appendix D.1, are measured at that level. For municipality m, we estimate the following equations:

$$B_m = \beta_e O_m^e + X_m^e \gamma_e + \varepsilon_m^e \quad (3)$$

$$O_m^\ell = \beta_\ell B_m + X_m^\ell \gamma_\ell + \varepsilon_m^\ell \quad (4)$$

where B_m is a binary variable equal to 1 if municipality m experienced a roadblock on 11/17, O_m^e and O_m^ℓ are measures of online mobilization in municipality m before and after 11/17, and X_m^e and X_m^ℓ are a set of controls. To construct O_m^e, we aggregate information on early online mobilization into a binary variable equal to 1 if the municipality belongs to the highest quartile in the number of Facebook groups created before 11/17 (including regional groups, apportioned by municipal population) and in the petition signature rate before 11/17. For O_m^ℓ, we use a binary variable equal to 1 if the municipality hosts a new Facebook group after 11/17, the log number of local Facebook groups created after 11/17 (including regional groups, apportioned by municipal population), the log number of members in these groups, and the log number of messages posted on these groups.

Instrumental variables. The OLS estimates of β_e and β_ℓ may suffer from several omitted variable biases, which go in opposite directions. First, online and offline mobilizations are both affected by unobservable characteristics such as latent discontent, which may induce an upward bias. Conversely, people living in different municipalities may unobservably vary in their preference for online vs. offline action, weakening the link between different protest stages. In addition, both independent and outcome

[25] For example, large street protests receive a lot of media coverage, and increase the salience of the mobilization. Consistent with this information channel, weekly street protests were associated with a sharp increase in Google queries about the Yellow Vests on Facebook (see Appendix Figure C.2).

variables are subject to measurement error stemming from our data collection process. In line with the counterfactual sequences depicted in Figure 3, we circumvent these issues by comparing localities where the cost of online and offline mobilization varies exogenously, using an instrumental variable strategy.

To show the causal impact of early online mobilization on street protests, we instrument O_m^e in Equation (3) with the presence of a 4G antenna in the municipality prior to 11/17, using the fact that the roll-out of 4G in France, albeit quite fast, was only about half complete at the time of the first protests.[26] Consistently with global evidence on the impact of cell phone access on the likelihood of protests (see, e.g. Pierskalla and Hollenbach, 2013; Manacorda and Tesei, 2020), access to 4G improves signal quality and thus the time people spend on their phones, which should increase the likelihood that they will hear about the petition or coordinate to form a local Facebook group. The identifying assumption behind this instrument is that, conditional on our extensive set of controls, the timing of the installation of 4G antennas was driven by operational constraints such as the date of frequency auctions or the availability of material and labor that were not correlated with unobserved drivers of discontent and mobilization (see Panels E and F of Appendix Figure D.1 for a map showing the seemingly random distribution of residualized 4G access).

In Equation (4), we instrument B_m with the presence of a highway toll in the municipality.[27] France has an extensive highway network spanning over 11,000 kilometers. Most of this network was constructed by the central government in the 1970s and 1980s to connect major cities to each other and to Paris. Seventy-five percent of the current network consists of toll roads.[28] The Yellow Vests targeted highway tolls as a symbol of the increased cost of driving. On 11/17, they took control of one hundred of them, where they organized slowdowns and allowed drivers to pass through for free. Indeed, we observe that 19% of municipalities with a toll were blocked, compared to 2% of other

[26] To retrieve information on the distribution of 4G Antennas, we use May 2024 official data from the *Agence Nationale des Fréquences*. These data show that 40,313 antennas were installed before 11/17 and 44,807 after. The roll-out of 4G was all but over in 2024, with close to complete coverage and the start of the 5G roll-out in 2020.

[27] An earlier version of this paper used the spatial distribution of road roundabouts as an instrument for roadblocks. The rationale for using this instrument is similar to that of tolls: roundabouts were heavily targeted by protesters, because they allowed them to block several roads and are easy to set camp on. However, roundabout locations appear to be correlated with past protest locations, unlike toll locations.

[28] See Panel D of Appendix Figure D.1 for a map. Free highways are scattered throughout France. Some are close to large cities or international borders, while others are on smaller, newer sections.

municipalities. After controlling for municipal characteristics and Living Zone fixed effects, this gap is halved but remains substantial. The identifying assumption we make is that the location of highway tolls, conditional on our extensive set of controls, was driven by operational constraints faced by the public planner decades ago and is not correlated with unobserved drivers of the mobilization at the very local level.[29]

To assess the validity of our instruments, we implement a placebo test exploiting the geography of demonstrations during the two most recent major episodes of social unrest before the Yellow Vests: the 2010 protests against the pension reform and the 2016 protests against the labor law, covering 13 protest days in total. The key threat to identification is that our instruments might capture durable local characteristics — e.g., social capital or latent discontent — that make certain places systematically prone to mobilization regardless of the specific movement. If this were the case, 4G access and toll presence should predict protest locations not only during the Yellow Vest episode but also during earlier, unrelated episodes. We test this directly by running a horse-race regression in which the dependent variable is the instrument and the regressors include indicators for having hosted a demonstration on each of these past protest days, together with an indicator for participation in the Yellow Vest episode, and our full set of controls and Living Zone fixed effects. Under the exclusion restriction, our instruments should capture factors specific to the Yellow Vest movement — the role of social media coordination for 4G, the symbolic and logistical salience of tolls for a movement centered on driving costs — rather than general protest propensity. Panel B of Appendix Figure D.2 reports the corresponding estimates. We find no systematic relationship between our instruments and the location of pre-Yellow Vest protests, and fail to reject the joint null hypothesis that all coefficients on past protest episodes equal zero (Toll: F=1.07, p=0.38; 4G: F=0.46, p=0.94). This supports the interpretation that our instruments operate through channels specific to the organizational features of the Yellow Vest movement.

3.2.2 Results

Our estimation results are presented in Columns (1) to (5) of Table 1, which show OLS and 2SLS estimates of β_e and β_ℓ (Panels A and C) as well as OLS estimates of

[29] See Faber (2014), among others, for a similar strategy. While we cannot test this assumption, we note that municipalities with tolls do not stand out in terms of observable characteristics, even when restricting the sample to municipalities located on highways: The adjusted R-squared of an OLS regression of the probability of hosting a toll on our set of controls is equal to 4%, compared to over 30% for the probability of a 11/17 roadblock. Therefore, the predictive power of tolls regarding blockades remains unchanged regardless of the number of control variables used (see Appendix Figure D.2).

the correlation between the instrument and the endogenous variable (Panel B). First, Column (1) shows estimates of Equation (3) on the impact of early online mobilization on the probability of a roadblock on 11/17. While the correlation between O^e_m and B_m is small (3 p.p.), the 2SLS estimate is much higher (24 p.p.), suggesting a strong positive impact of early online mobilization on the organization of the 11/17 protests. This result is not surprising given many Facebook groups were actually created with the explicit purpose of facilitating the organization of the upcoming blockades, and it is consistent with a vast body of evidence in other contexts.

We then turn to our estimates of β^ℓ on the impact of the 11/17 blockades on subsequent online mobilization. Column (2) shows that a blockade on 11/17 increases the likelihood of the creation of a new Facebook group in the municipality by 44 p.p., which is, once again, a very large effect, albeit consistent with the surge displayed in Panel C of Figure 5. Columns (3) to (5) show that the causal effects on other measures of Facebook activity are also very large. A roadblock increases the log number of new local groups by 1.13 (Column 3). Moreover, these groups are not small, nor inactive: the effects of a blockade on the log number of members in these new local groups (Column 4) and the log number of messages posted on these groups (Column 5) are equal to 1.31 and 1.26, respectively.

The 2SLS estimates are larger than their OLS counterparts across all specifications. A natural explanation is that both endogenous regressors are measured with error, which attenuates OLS estimates toward zero. In Equation (3), early online mobilization O^e_m is constructed from Facebook groups and petition signatures, but other forms of online coordination — such as private messaging or activity on platforms we do not observe — may also have contributed to the organization of the 11/17 protests. In Equation (4), the blockade indicator B_m captures roadblocks announced on a coordination website, potentially missing spontaneous events. In both cases, classical measurement error in the regressor generates attenuation bias that valid instruments correct. While we cannot test for the presence of attenuation bias in the case of O^e_m, we show in Appendix Table D.1 that restricting the set of blocked municipalities to those where a newspaper report of the protest is available would further increase the absolute gap between the OLS and 2SLS estimates of β^ℓ. Conversely, as expected, using this more conservative definition of B_m yields lower estimates of β^e.

Robustness checks. Appendix Tables D.2 to D.6 show the robustness of these results to several concerns. First, both F-statistics and 2SLS estimates are remarkably stable in specifications with fewer covariates. In fact, a very parsimonious set of baseline controls

Table 1: Feedback Loop Between Online and Offline Mobilization

	Dependent Variable:				
	Offline 11/17	Online Mobilization on Facebook Post-11/17			
	Blockade (indicator)	Group (indicator)	Groups (in logs)	Members (in logs)	Posts (in logs)
	(1)	(2)	(3)	(4)	(5)
Panel A: OLS					
Online Pre-11/17	0.034***				
	(0.006)				
Blockade		0.114***	0.295***	0.271***	0.265***
		(0.016)	(0.046)	(0.057)	(0.059)
Panel B: Reduced form					
4G Antenna	0.006***				
	(0.002)				
Toll		0.044***	0.113***	0.131***	0.126***
		(0.012)	(0.038)	(0.048)	(0.049)
Panel C: 2SLS					
Online Pre-11/17	0.242***				
	(0.092)				
Blockade		0.445***	1.127***	1.312***	1.261**
		(0.124)	(0.396)	(0.490)	(0.492)
Controls	✓	✓	✓	✓	✓
Mean dep. var.	0.02	0.02	-5.15	-0.01	-0.13
Observations	34475	34475	34475	34475	34475
Robust F Stat	28.38	48.34	48.34	48.34	48.34

Notes: The outcome and explanatory variables are described in the text. All explanatory variables are binary. Panel A shows the OLS estimates of the correlation between O^e and B (Column 1) and between B and O^ℓ (Columns 2 to 5). Panel B shows the OLS estimates of the correlation between our outcome variables and our instruments: a binary variable equal to 1 if the municipality has a working 4G antenna before 11/17 and a binary variable equal to 1 if the municipality hosts a highway toll. Panel C shows the corresponding 2SLS estimates. Controls are listed in Appendix D.1 and include a set of over 1,600 Living Zone fixed effects. We cluster standard errors at the Living Zone level. *: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$.

made of the log of municipal population and three indicators that the municipality is an administrative center at the county, district or subdistrict level delivers very similar results. Controlling for various proxies of commuting cost, which was the initial trigger of the petition, for political preferences with past election results, or for regional differ-

ences with an extensive set of Living Zone fixed effects does not affect the estimates. Controlling, in the case of Equation (4), for our measure of early online mobilization (and the presence of a 4G antenna in the municipality) to allow for possible substitution effects between different phases of online mobilization is also inconsequential. Finally, the results are also remarkably stable if we restrict the sample to municipalities outside the Paris region, which stands alone along many dimensions, or, in the regressions using the highway toll variable, to the small subset of municipalities that are located on a highway and may therefore be more comparable to the municipalities with a highway toll.

In a second series of robustness checks, we use alternative independent variables and instruments. First, we assess how sensitive our estimates of β^e are when using alternative cutoff quantiles to define O_m^e. As expected, Appendix Table D.7 shows that the OLS correlations and 2SLS estimates increase with more restrictive definitions. Compared to our baseline estimate (based on quartiles), the estimate for β^e is twice as high when using octiles and nearly four times higher when using deciles. However, the F-statistic is lower with more restrictive definitions that exclude too many municipalities with 4G coverage. Finally, we use a second instrument for the 11/17 blockades to test overidentifying restrictions. Since organizing a roadblock requires significant manpower, protesters must have coordinated to choose roadblock locations. This spatial coordination problem suggests another instrument: the presence of a toll in one of the other municipalities in the Living Zone, which is the mirror image of the first instrument. Due to competition among easily blockable locations, we expect municipalities not surrounded by tolls to be more likely blocked. Appendix Table D.8 shows that the effects of the instruments go in the expected directions. The F-statistic equals 35 when using the second instrument only and 25 when using both instruments simultaneously. Additionally, the high p-values associated with the Hansen J-statistics indicate that we cannot reject the hypothesis that the overidentifying restrictions are valid. The resulting 2SLS estimates are fairly comparable to those in Table 1.

3.3 Crowding out: online violence and the departure of the moderates

We now present evidence consistent with the crowd-out sequence depicted in Figure 4. Specifically, we show that (i) the surge of online mobilization following 11/17 was also a surge of online violence; (ii) the radicalization of the movement involved both an intensive margin (the average protester became more violent) and an extensive margin (radical protesters replaced moderate protesters); (iii) moderates left partly because of

others' radicalization. Ideally, points (ii) and (iii) should be assessed both online and offline. However, we do not have access to panel data on street protesters.[30] Moreover, participation to street protests after 11/17 was partly determined by police response, on which we have no reliable information. For these reasons, we focus on online mobilization, which provides a more controlled environment.

3.3.1 The revelation of a higher share of radicals

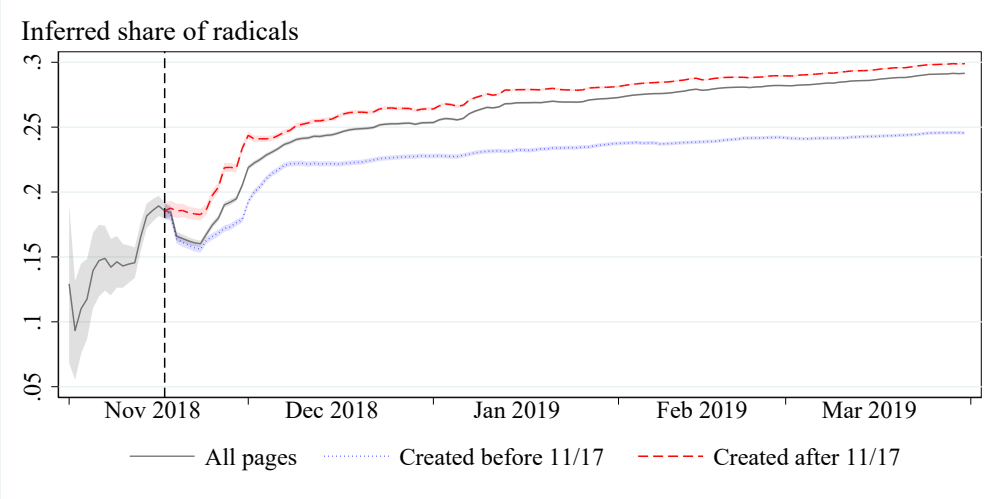
Since we lack information on the content posted on Facebook groups, we turn to our second dataset, which describes discussions on public Facebook pages. As shown in Panel D of Figure 5, many of those pages were created just after 11/17, mirroring our 2SLS results on the formation of new groups after the first protests. The creation of these new pages gives us the opportunity to conduct the following thought experiment and ask: if the Yellow Vest movement had not enriched its online infrastructure with new pages after the 11/17 protests, would this have changed protesters' beliefs about the share of radicals in the movement?

To mimic the belief updating process described in our theoretical framework, we construct an "inferred share of radicals" as the cumulated share of antagonistic sentences posted on Facebook pages since the beginning of November. Figure 7 shows that this inferred share increases over time, consistently with Panel C in Figure 6. However, it also shows that half of the increase (from 20% on 11/17 to 30% at the end of our study period) was driven by pages created after 11/17, where discussions were more radical from the start. Within two days, estimates derived from newer and older pages differ by 3 p.p., and they keep diverging afterwards.[31] These results suggest that the new wave of online mobilization following the 11/17 protests, which our 2SLS estimates show was at least partly caused by those protests, led to more violent expression and to a sudden upward shift in the perceived share of radicals in the population of protesters, in line with Proposition 3.

[30] This type of data may become available in the future, for example in high-tech autocracies using facial recognition. However, besides ethical concerns, this data will never be accessible to researchers.

[31] See Appendix Figure E.8 for the same analysis on our two other measures of online radicalization. To investigate whether this pattern was indeed specific to pages created in the aftermath of the 11/17 protests, we replicate this exercise with four other creation dates. Appendix Figure E.9 shows that pages created at different dates do not show a surge of radical activity around their creation date: in fact, they tend to be slightly less radical, before catching up with earlier pages. This is even the case for pages created after the most violent day of protests (December 1st).

Figure 7: Bayesian updating on the share of radical discussants



Notes: This figure shows the updating of the share of radicals on Facebook pages, based on the fraction of sentences that belong to an antagonistic topic. The update is based on observations from the previous days and the 95% confidence interval is computed with the binomial distribution. In black, the update is conducted using all available information. In blue, the update after 11/17 is conducted using information from pages created before 11/17. In red, the update after 11/17 is conducted using information from pages created on or after 11/17. The vertical dashed line corresponds to 11/17.

3.3.2 The two margins of online radicalization

After this sudden inflow of radical content from pages created after 11/17, discussions kept radicalizing over the following months. We now turn to the decomposition of this radicalization process between an extensive margin and an intensive margin. The intensive margin of radicalization measures whether the average user has become increasingly likely to post radical messages. As for the extensive margin, it measures whether the pool of active users has become increasingly populated with users who (on average) post more radical messages. Anecdotally, we can observe these two margins in our text corpus, as the tension between moderates and radicals unfolds on Facebook pages.[32] To disentangle both margins, we use a subsample of our data with user identi-

[32]For instance, a protester condemns street violence and worries it will discredit the movement: *"People are surprised to see Emmanuel Macron's rise in the polls... Could we reasonably think that the initial popular support would last forever in the current context? I mean, in a context of recurring violence."* Another protester wonders about their own participation in street protests that are expected to be violent: *"I went to protest for the first time in Bordeaux with the Yellow Vests. I arrived a little anxious and despairing and afraid of the violence of the excesses."* In line with the role played by the intensive margin, many protesters progressively become more radical over time. In November, a protester writes: *"Bravo to all of you, you are amazing."* as well as *"Bravo to you, gentlemen police officers, for*

fier to estimate the following equation:

$$Y_{s,i,t} = \delta_i + \gamma_t + \varepsilon_s, \tag{5}$$

where $Y_{s,i,t}$ is a measure of radicalism of sentence s written by user i in month t, δ_i is a user fixed effect, and γ_t is a month fixed effect. Intuitively, δ_i measures user i's propensity to post radical sentences, and γ_t accounts for the additional propensity of users to post radical sentences during month t.

We can then leverage estimates of user and time fixed effects to decompose the rise of online radicalism into an intensive and extensive margin. Indeed, the average level of radical sentences during month t, $\mathbb{E}_t[Y]$, can be expressed as:

$$\mathbb{E}_t[Y] = \underbrace{\mathbb{E}_t[\delta]}_{\text{Extensive margin}} + \underbrace{\gamma_t}_{\text{Intensive margin}}, \tag{6}$$

where $\mathbb{E}_t[\delta] = \sum_i s_{i,t}\delta_i$ and $s_{i,t}$ is the share of sentences posted during month t that originated from user i. Hence, the first term of Equation (6) corresponds to the average propensity to post radical sentences for users active during month t. An increase of this term over time means that the share of sentences posted by more radical users increases. An increase in the second term of Equation (6) corresponds to an increase in the propensity of any given user to post a radical sentence at a given time.

Panel A of Figure 8 presents a decomposition of our radicalization measures using the empirical counterpart of Equation (6). This decomposition suggests that both margins contributed almost equally to the radicalization of Facebook content.[33] Moreover, the effect of the extensive margin appears to be slightly delayed relative to that of the intensive margin, suggesting that the radicalization of some discussants triggered the defection of the more moderate ones.

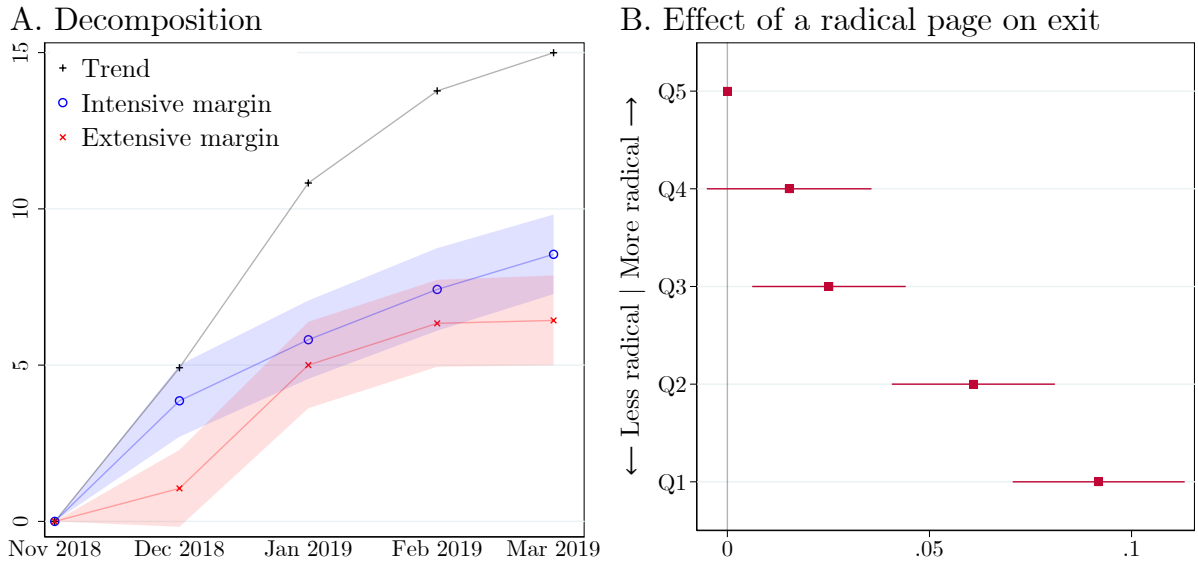
3.3.3 The crowding out of moderates

While this decomposition result is compatible with the crowding out of moderate Yellow Vest supporters, other mechanisms were plausibly at play. In December 2018, the government abandoned the planned gas tax hike and subsequently announced a

your support. You are courageous." Yet, in December and January, their tone markedly changes with messages such as: *"Reduce these ******* to nothing."* and *"All corrupt, these *******."*

[33] See panels A1 and B1 of Appendix Figure E.10 for the same analysis on our two other proxies for radicalism.

Figure 8: The crowding out of moderate online protesters



Notes: Panel A shows the decomposition of the increase in online radicalism (in %) using Equation (6). We compute standard errors via bootstrap with 1000 iterations and plot 95% confidence intervals. Panel B shows the OLS estimates of β_1, β_2, β_3 and β_4 of Equation (7) and their associated 95% confidence intervals, with standard errors clustered at the discussant level. β_5 is set to zero by normalization. In both panels, radicalism is measured at the sentence level as a binary variable equal to 1 if the sentence belongs to an antagonistic topic.

generous income redistribution package. Moreover, some street protests were met with heavy-handed policing, and many online discussions mention incidents with the police. This dual response may have simultaneously reduced the incentives for more moderate protesters to participate and antagonized more radical protesters. In order to directly test for the crowding-out mechanism highlighted in point (iii) in Proposition 3, we use the previous empirical framework to measure the impact of discussion radicalization on the online mobilization of different types of protesters, while controlling for time-specific confounders. For each measure of radicalism, we first estimate discussant fixed effects using Equation (5). Then, on the sample of discussant-by-page-by-month observations, we estimate the following equation:

$$\mathbb{P}(\text{Exit})_{i,p,t} = \sum_q \beta_q \left(\mathbb{1}_{\delta_i \in q} \times \mathbb{E}_{p,t}\left[\delta\right] \right) + \zeta_i + \xi_{p,t} + \boldsymbol{X}_{i,p,t}\eta + \varepsilon_{i,p,t}, \tag{7}$$

where $\mathbb{P}(\text{Exit})_{i,p,t}$ is a binary variable indicating that discussant i stops posting on page p after month t, $\mathbb{1}_{\delta_i \in q}$ is a binary variable indicating to which quantile (evaluated over the population of discussants) the discussant's radicalism fixed effect belongs, $\mathbb{E}_{p,t}\left[\delta\right]$ is the (standardized) average of the discussant radicalism fixed effect associated with

sentences posted on page p during month t, ζ_i is a discussant dummy, $\xi_{p,t}$ is a page-by-month dummy, and $X_{i,p,t}$ is a vector of additional controls at the discussant-by-page-by-month level.[34] Importantly, this specification controls for potential aggregate changes in the cost of participation over time, which could be driven by protester fatigue or online repression and would differentially affect moderates and radicals.

Our results are summarized in Panel B of Figure 8, which breaks down individual radicalism into quintiles. These results fully support the hypothesis that more radical discussants crowded out moderate ones. For a discussant whose fixed effect belongs to the first quintile of radicalism (the least radical), being exposed to a page where the average level of discussant radicalism is one standard deviation above the mean increases their probability to stop posting on that page by 9 p.p., or 14% of the baseline probability, compared to a discussant in the fifth quantile of radicalism. This effect decreases monotonically with the level of individual radicalism and is not statistically different from zero for the more radical half of the discussants.[35]

Robustness checks and discussion. We evaluate the robustness of this result along several dimensions. First, one may consider that a better measure of page radicalism would be the radicalism of the average posted sentence ($\mathbb{E}_{p,t}[Y]$), rather than the average value of discussants' radicalism fixed effects ($\mathbb{E}_{p,t}[\delta]$). While this measure, computed on more observations, is less subject to measurement error, it may also be polluted by period-specific effects that are accounted for in our first stage. However, as shown in Panel A1 of Appendix Figure E.11, the results are remarkably similar if we use this alternative measure of page radicalism. Similarly, the results are robust to computing page radicalism without including the sentences posted by the discussant themselves (See Panel A2).[36]

[34] For this second stage, we restrict the estimation sample to pages that are still active the following months. The sample comprises 67,957 user-page-month observations, for 24,076 users and 292 pages (See Appendix Table E.7 for details). The distribution of activity is very skewed: 62% of users post twice, 21% post three times, and 1% of users post eight times or more. We control for the number of sentences posted by the discussant during month t, either on page p or on other pages. We also control for a binary variable indicating whether the discussant had already posted on the page before month t. In practice, excluding these controls is inconsequential. For the estimation, we replace expectations and quantiles of δ_i by their empirical counterparts using our estimates of Equation (5).

[35] See Panels A2 and B2 of Appendix Figure E.10 for the same analysis on our two other proxies for radicalism.

[36] We also tested a specification without prior estimation of individual fixed effects, defining radicalism of discussant i at month t as the share of radical comments posted

Second, we show that our result is not driven by spurious correlation due to an overly saturated model. While we believe that the best specification should include discussant and page-by-month fixed effects to control for discussants sorting across pages and the unobservable time-varying characteristics of each page, we replicate the analysis with a less restrictive set of fixed effects. Our results are reported in Appendix Table E.7. The coefficients associated with our variable of interest are all positive and statistically significant. Moreover, they tend to increase with the richness of the set of fixed effects, which suggests that moderate discussants sort across pages according to their tolerance for radical discussion, even if they do not post radical messages themselves.

Third, the average crowding out effect we measure may mask substantial variation over our study period. On the one hand, tolerance for radical discussion may have increased over time due to the individual radicalization process depicted in Panel A of Figure 8 and the associated shift in norms regarding what is considered acceptable in a conversation. This effect would bias our estimates downward. On the other hand, decisions to leave a page may reflect the entire history of discussants: for example, they may decide to leave a page only after they have reached their maximum cumulative level of exposure to radical content over time. In this case, our estimates would also capture this tipping mechanism and could be biased upward. However, consistently with our modeling choice to consider myopic players, our results suggest that these dynamic concerns are not of first order. As shown in column (5) of Appendix Table E.7, estimates are remarkably stable when we control for discussant-by-month fixed effects, which can be estimated for the subset of discussants who post simultaneously on multiple pages during the month.

Fourth, to check whether the crowding-out effect we observe is specific to the decision to leave the focal page, we replicate the analysis on the probability of leaving any other page where the discussant is also active. Results shown in Panel B1 of Appendix Figure E.11 confirm that crowding out is specific to the focal page: moderates are not more likely to leave other pages when exposed to radical content on a given page. In fact, they become slightly less likely to leave the other pages. However, this indirect positive effect is twice lower in magnitude than the direct negative effect, so that, on average, moderates are still 5 p.p. more likely to exit at least one of the pages where they currently post when they are exposed to radical content on one of those pages (see Panel B2).

by the discussant in the months preceding month t. In practice, we replace δ_i by $\mathbb{E}_{i,t'<t}[Y]$ in Equation (7) and we measure quantiles q on the population of discussants observed during month t. The results are very similar.

Finally, our methodology is based on the assumption that the unweighted average of all sentences is an accurate representation of exposure to radical content. This assumption conflicts with the well-documented tendency of social media platforms to highlight antagonistic content. In Appendix E.7, we provide evidence suggesting that Facebook's recommendation algorithm made radical statements more visible to the average user. This feature may have increased overall exposure to radical content, and precipitated the departure of moderate discussants, as discussed in Section 2.3. However, platforms also try to cater to individual needs by personalizing the user experience (Matter and Hodler, 2025), and moderate discussants may have actually been less exposed to radical content than radical users. Testing these two factors would require access to individual browser histories.

4 Conclusion

Protest movements seek to form large coalitions, but these coalitions are susceptible to fracture when protests turn violent. This paper examines this tension, which has been at the heart of many episodes of social unrest. To do so, it draws on a salient feature of contemporary protest movements: their use of social media. We propose a conceptual framework in which social media can both increase the likelihood of protests and increase the likelihood that initially successful protest movements will eventually turn violent and fade away. We then show that the theoretical mechanisms we highlight are consistent with the history of the Yellow Vest movement.

We view our results as a cautionary tale about the impact of social media on the effectiveness of protest movements. When protest movements seek only to organize one-off events (e.g., to raise awareness about a particular issue), social media may prove effective by helping to mobilize a higher proportion of the population; conversely, when protest movements need to wage protracted campaigns to achieve their goals (e.g., to force substantial policy changes on the government), social media may prove detrimental by revealing to the coalitions behind the movement how heterogeneous they are, which may convince different factions to adopt divergent and possibly mutually exclusive mobilization strategies.

Our analysis abstracts from other plausible mechanisms. In particular, we believe that the process of gradual revelation we propose is more general than our application: for example, beyond protest tactics, protesters may also come to realize that they do not share the same goals with each other. Collecting data on different aspects of protesters' beliefs in real time would help to disentangle these mechanisms.

References

Acemoglu, Daron, Tarek A Hassan, and Ahmed Tahoun, "The Power of the Street: Evidence From Egypt's Arab Spring," *Review of Financial Studies*, 2018, *31* (1), 1–42.

Algan, Yann, Elizabeth Beasley, Daniel Cohen, Martial Foucault, and Madeleine Péron, "Qui Sont Les Gilets Jaunes Et Leurs Soutiens," Technical Report, CEPREMAP and CEVIPOF 2019.

Ananyev, Maxim, Dimitrios Xefteris, Galina Zudenkova, and Maria Petrova, "Information and Communication Technologies, Protests, and Censorship," 2019. Working Paper.

Angeletos, George-Marios, Christian Hellwig, and Alessandro Pavan, "Dynamic global games of regime change: Learning, multiplicity, and the timing of attacks," *Econometrica*, 2007, *75* (3), 711–756.

Aridor, Guy, Rafael Jiménez-Durán, Ro'ee Levy, and Lena Song, "The Economics of Social Media," *Journal of Economic Literature*, December 2024, *62* (4), 1422â74.

Ash, Elliott and Stephen Hansen, "Text Algorithms in Economics," *Annual Review of Economics*, 2023, *15* (Volume 15, 2023), 659–688.

Aumann, Robert, "Acceptable points in General Cooperative *n*-person Games," in Albert William Tucker and Robert Duncan Luce, eds., *Contributions to the Theory of Games (AM-40), Volume IV*, Princeton University Press, 1959, pp. 287–324.

Bail, Christopher A., Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky, "Exposure to opposing views on social media can increase political polarization," *Proceedings of the National Academy of Sciences of the United States of America*, 2018, *115* (37), 9216–9221.

Barbera, Salvador and Matthew O. Jackson, "A Model of Protests, Revolution, and Information," *Quarterly Journal of Political Science*, July 2020, *15* (3), 297–335.

Battaglini, Marco, "Public Protests and Policy Making," *Quarterly Journal of Economics*, 2017, *132* (1), 485–549.

_ , Rebecca Morton, and Eleonora Patacchini, "Social Groups and the Effectiveness of Petitions," Technical Report, NBER Working Paper 26757 2020.

Bohren, J Aislinn and Daniel N Hauser, "Learning with heterogeneous misspecified models: Characterization and robustness," *Econometrica*, 2021, *89* (6), 3025–3077.

Boyer, Pierre C., Thomas Delemotte, Germain Gauthier, Vincent Rollet, and Benoît Schmutz, "The Origins of the Gilets Jaunes Movement," *Revue Économique*, 2020, *71* (1), 109–138.

Brummitt, Charles D., George Barnett, and Raissa M. D'Souza, "Coupled catastrophes: sudden shifts cascade and hop among interdependent systems," *Journal of The Royal Society Interface*, 2015, *12* (112), 20150712.

Bueno de Mesquita, Ethan, "Rebel Tactics," *Journal of Political Economy*, 2013, *121* (2), 323–357.

Bueno de Mesquita, Ethans, "Regime Change and Revolutionary Entrepreneurs," *American Political Science Review*, 2010, *104* (3), 446–466.

Bursztyn, Leonardo, Davide Cantoni, David Yang, Noam Yuchtman, and Jane Zhang, "Persistent Political Engagement: Social Interactions and the Dynamics of Protest Movements," *American Economic Review: Insights*, 2021, *3* (2), 233–50.

_ , Georgy Egorov, Ruben Enikolopov, and Maria Petrova, "Social Media and Xenophobia: Evidence from Russia," 2024. Working paper.

Cantoni, Davide, Andrew Kao, David Y. Yang, and Noam Yuchtman, "Protests," *Annual Review of Economics*, 2024, pp. 519–543.

_ , David Y. Yang, Noam Yuchtman, and Y. Jane Zhang, "Protests as Strategic Games: Experimental Evidence From Hong Kong's Antiauthoritarian Movement," *Quarterly Journal of Economics*, 01 2019, *134* (2), 1021–1077.

Casanueva, Annalí, "Can Chants in the Street Change Parliament's Tune? The Effects of the 15M Social Movement on Spanish Elections," 2025. Working paper.

Clarke, Killian and Korhan Kocak, "Launching Revolution: Social Media and the Egyptian Uprising's First Movers," *British Journal of Political Science*, 2020, *50* (3), 1025–1045.

Correa, Sofia, "Persistent Protests," *American Economic Journal: Microeconomics*, May 2025, *17* (2), 321–57.

Demszky, Dorottya, Nikhil Garg, Rob Voigt, James Zou, Matthew Gentzkow, Jesse Shapiro, and Dan Jurafsky, "Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings," in "Proceedings of NAACL-HLT" 2019, pp. 2970–3005.

Douenne, Thomas and Adrien Fabre, "Yellow Vests, Pessimistic Beliefs, and Carbon Tax Aversion," *American Economic Journal: Economic Policy*, 2022, *14* (1), 81–110.

Enikolopov, Ruben, Alexey Makarin, and Maria Petrova, "Social Media and Protest Participation: Evidence From Russia," *Econometrica*, 2020, *88* (4), 1479–1514.

_ , _ , _ , and Leonid Polishchuk, "Social Image, Networks, and Protest Participation," 2020. Working paper.

Esponda, Ignacio and Demian Pouzo, "Berk–Nash equilibrium: A framework for modeling agents with misspecified models," *Econometrica*, 2016, *84* (3), 1093–1130.

Faber, Benjamin, "Trade Integration, Market Size, and Industrialization: Evidence from China's National Trunk Highway System," *The Review of Economic Studies*, 03 2014, *81* (3), 1046–1070.

Fergusson, Leopoldo and Carlos Molina, "Facebook Causes Protests," 2021. Working paper.

Fudenberg, Drew and David K. Levine, "Self-Confirming Equilibrium," *Econometrica*, 1993, *61* (3), 523–545.

Fujiwara, Thomas, Karsten Muller, and Carlo Schwarz, "The Effect of Social Media on Elections: Evidence from The United States," *Journal of the European Economic Association*, 10 2024, *22* (3), 1495–1539.

Gaby, Sarah and Neal Caren, "Occupy Online: How Cute Old Men and Malcolm X Recruited 400,000 US Users to OWS on Facebook," *Social Movement Studies*, 2012, *11* (3-4), 367–374.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, "Text as Data," *Journal of Economic Literature*, 2019, *57* (3), 535–74.

Gethin, Amory and Vincent Pons, "Social Movements and Public Opinion in the United States," Working Paper 32342, National Bureau of Economic Research April 2024.

Gieczewski, Germán and Korhan Kocak, "Collective procrastination and protest cycles," *American Journal of Political Science*, 2024, *n/a* (n/a).

Granovetter, Mark, "Threshold Models of Collective Behavior," *American Journal of Sociology*, 1978, *83* (6), 1420–1443.

Grimmer, Justin and Brandon M Stewart, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis*, 2013, *21* (3), 267–297.

Gylfason, Gisli, "From Tweets to the Streets: Twitter and Extremist Protests in the United States," PSE Working Papers halshs-04188189, HAL 2025.

Hager, Anselm, Lukas Hensel, Johannes Hermle, and Christopher Roth, "Group size and protest mobilization across movements and countermovements," *American Political Science Review*, 2022, *116* (3), 1051–1066.

Hassan, Mai, "Coordinated Dis-Coordination," *American Political Science Review*, 2021, pp. 1–15.

Kricheli, Ruth, Yair Livne, and Beatriz Magaloni, "Taking to the Streets: Theory and Evidence on Protests under Authoritarianism," 2011. Working paper.

Larson, Jennifer M, Jonathan Nagler, Jonathan Ronen, and Joshua A Tucker, "Social Networks and Protest Participation: Evidence From 130 Million Twitter Users," *American Journal of Political Science*, 2019, *63* (3), 690–705.

Levy, Ro'ee, "Social Media, News Consumption, and Polarization: Evidence from a Field Experiment," *American Economic Review*, 2021, *111* (3), 831–870.

Little, Andrew T., "Communication Technology and Protest," *Journal of Politics*, 2016, *78* (1), 152–166.

Little, Andrew T, "Coordination, Learning, and Coups," *Journal of Conflict Resolution*, 2017, *61* (1), 204–234.

Loeper, Antoine, Jakub Steiner, and Colin Stewart, "Influential Opinion Leaders," *Economic Journal*, 2014, *124* (581), 1147–1167.

Lohmann, Susanne, "A Signaling Model of Informative and Manipulative Political Action," *American Political Science Review*, 1993, *87* (2), 319–333.

Manacorda, Marco and Andrea Tesei, "Liberation Technology: Mobile Phones and Political Mobilization in Africa," *Econometrica*, 2020, *88* (2), 533–567.

Matter, Ulrich and Roland Hodler, "Web Search Personalization During the US 2020 Election," *American Economic Review: Insights*, 2025, *7* (4), 516–33.

Morris, Stephen and Hyun Song Shin, "Unique equilibrium in a model of self-fulfilling currency attacks," *American Economic Review*, 1998, pp. 587–597.

_ **and Mehdi Shadmehr**, "Inspiring Regime Change," *Journal of the European Economic Association*, 2023, *21* (6), 2635–2681.

_ **and** _ , "Repression and repertoires," *American Economic Review: Insights*, 2024, *6* (3), 413–433.

Petrovskii, Sergei, Maxim Shishlenin, and Anton Glukhov, "Understanding street protests: from a mathematical model to protest management," *PLOS ONE*, 04 2025, *20*.

Pierskalla, Jan H. and Florian M. Hollenbach, "Technology and Collective Action: The Effect of Cell Phone Coverage on Political Violence in Africa," *The American Political Science Review*, 2013, *107* (2), 207–224.

Rieder, Bernhard, "Studying Facebook via Data Extraction: The Netvizz Application," in "Proceedings of the 5th annual ACM web science conference" ACM 2013, pp. 346–355.

Ross-Arguedas, Amy, Craig Robertson, Richard Fletcher, and Rasmus Nielsen, "Echo Chambers, Filter Bubbles, and Polarisation: A Literature Review," Technical Report, Reuters Institute for the Study of Journalism 2022.

Rydzak, Jan, Moses Karanja, and Nicholas Opiyo, "Internet Shutdowns in Africa—Dissent Does Not Die in Darkness: Network Shutdowns and Collective Action in African Countries," *International Journal of Communication*, 2020, *14* (0).

Shadmehr, Mehdi and Dan Bernhardt, "Collective Action with Uncertain Payoffs: Coordination, Public Signals, and Punishment Dilemmas," *American Political Science Review*, 2011, *105* (4), 829–851.

_ **and** _ , "State censorship," *American Economic Journal: Microeconomics*, 2015, *7* (2), 280–307.

Shultziner, Doron and Irit Kornblit, "French Yellow Vests (Gilets Jaunes): Similarities and Differences With Occupy Movements," *Sociological Forum*, 02 2020, *35*.

Steinert-Threlkeld, Zachary C, "Spontaneous collective action: Peripheral mobilization during the Arab Spring," *American Political Science Review*, 2017, *111* (2), 379–403.

_ , Alexander M Chan, and Jungseock Joo, "How State and Protester Violence Affect Protest Dynamics," *Journal of Politics*, 2022, *84* (2), 798–813.

Tufekci, Zeynep, *Twitter and Tear Gas: The Power and Fragility of Networked Protest*, New Haven; London: Yale University Press, 2017.

Winters, Matthew S. and Rebecca Weitz-Shapiro, "Partisan Protesters and Nonpartisan Protests in Brazil," *Journal of Politics in Latin America*, 2014, *6* (1), 137–150.

Yao, Elaine, "Protest tactics and organizational structure," 2024. Working paper.

Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov, "Political Effects of the Internet and Social Media," *Annual Review of Economics*, 2020, *12*, 415–438.

Appendix

A Proofs

A.1 Analysis of the Stage Game

We start by stating the condition required for each of the five types of equilibrium to exist for some parameter region. Let $\underline{\theta}_R$ and $\overline{\theta}_R$ be defined as in Lemma 1. We assume that

Assumption 1

$$\underline{\theta}_R < \frac{\overline{c}}{v+1} < \overline{\theta}_R.$$

Proof of Lemma 1 The profile $(\mathbf{P}, \mathbf{V})$ is an equilibrium if and only if

$$\begin{cases} \theta_M + \alpha \mathbb{E}[\mu] - (\alpha + \beta)\mathbb{E}[\lambda\mu] \geq \underline{c}, \\ v\theta_R + (\gamma - \alpha)\mathbb{E}[\lambda\mu] \geq \overline{c} - \underline{c}, \end{cases}$$

i.e., if and only if $\theta_M \geq \overline{\theta}_M$ and $\theta_R \geq \overline{\theta}_R$. The first inequality reflects that moderates (collectively) prefer playing $a_M = \mathbf{P}$ to playing $a_M = \mathbf{A}$, and the second that radicals prefer playing $a_R = \mathbf{V}$ to $a_R = \mathbf{P}$. Note that these conditions also imply that no individual deviation is profitable. Assumption 1 also rules out a deviation by radicals to playing $a_R = \mathbf{A}$, since $\theta_R \geq \overline{\theta}_R$ implies $(v+1)\theta_R + \alpha(1-\lambda)\mu + \gamma\lambda\mu - \overline{c} \geq (v+1)\overline{\theta}_R - \overline{c} > 0$.

Similarly, the profile $(\mathbf{P}, \mathbf{P})$ is an equilibrium if and only if

$$\begin{cases} \theta_M + \alpha \mathbb{E}[\mu] \geq \underline{c}, \\ v\theta_R + (\gamma - \alpha)\mathbb{E}[\lambda\mu] \leq \overline{c} - \underline{c}, \end{cases}$$

i.e. if and only if $\theta_M \geq \underline{\theta}_M$ and $\theta_R \leq \overline{\theta}_R$. The first condition rules out a deviation to playing $a_M = \mathbf{A}$ by moderates, while the second condition rules out a deviation by radicals to playing $a_R = \mathbf{V}$. Last, under these conditions we have $\theta_R + \alpha\mathbb{E}[\mu] \geq \underline{c}$ as well, which guarantees that radicals do not prefer playing $a_R = \mathbf{A}$.

Similarly (we now omit details), the profile $(\mathbf{A}, \mathbf{V})$ is an equilibrium if and only if $\theta_M \leq \overline{\theta}_M$ and $\theta_R \geq \overline{\theta}_R$. The profile $(\mathbf{A}, \mathbf{P})$ is an equilibrium if and only if $\theta_M \leq \underline{\theta}_M$ and $\theta_R \in [\underline{\theta}_R, \overline{\theta}_R]$. Last, the profile $(\mathbf{A}, \mathbf{A})$ is an equilibrium if and only if $\theta_R \leq \underline{\theta}_R$ and $\theta_M + \alpha\mathbb{E}[(1-\lambda)\mu] \leq \underline{c}$. The condition $\theta_R \leq \underline{\theta}_R$ also implies that $(v+1)\theta_R < \underline{c}$, and thus radicals do not want to deviate and play $a_R = \mathbf{V}$.

Examining the regions of the plan (θ_M, θ_R) that sustain each type of equilibrium reveals that an equilibrium always exists, and it is unique except:

1. in the region where $v\theta_R + (\gamma - \alpha)\mathbb{E}[\lambda\mu] \leq \overline{c} - \underline{c}$ and $\theta_M + \alpha\mathbb{E}[(1-\lambda)\mu] \leq \underline{c} \leq \theta_M + \alpha\mathbb{E}[\mu]$ where $(\mathbf{A}, \mathbf{A})$ and $(\mathbf{P}, \mathbf{P})$ coexist; in that region, our second refinement criterion selects $(\mathbf{P}, \mathbf{P})$, which Pareto-dominates $(\mathbf{A}, \mathbf{A})$;

2. on knife-edge cases corresponding to the frontiers of the regions on Figure 1, we apply a tie-breaking criterion, which orders equilibria as follows: $(\mathbf{P}, \mathbf{P}) \succ (\mathbf{P}, \mathbf{V}) \succ (\mathbf{A}, \mathbf{A}) \succ (\mathbf{A}, \mathbf{P}) \succ (\mathbf{A}, \mathbf{V})$. This order is arbitrary and is only meant to facilitate the statements of formal results, e.g. Lemma 2, but none of our qualitative results relies on it.

To summarize, the equilibrium is unique for all parameter values, and characterized as follows:

1. if $\theta_R < \underline{\theta}_R$ and $\theta_M < \underline{\theta}_M$, the equilibrium is $(\mathbf{A}, \mathbf{A})$;

2. if $\underline{\theta}_R \leq \theta_R \leq \overline{\theta}_R$ and $\theta_M < \underline{\theta}_M$, the equilibrium is $(\mathbf{A}, \mathbf{P})$;

3. if $\overline{\theta}_R < \theta_R$ and $\theta_M < \overline{\theta}_M$, the equilibrium is $(\mathbf{A}, \mathbf{V})$;

4. if $\theta_R \leq \overline{\theta}_R$ and $\theta_M \geq \underline{\theta}_M$, the equilibrium is $(\mathbf{P}, \mathbf{P})$;

5. if $\overline{\theta}_R < \theta_R$ and $\overline{\theta}_M \leq \theta_M$, the equilibrium is $(\mathbf{P}, \mathbf{V})$.

A.2 Proof of Lemma 2

Fix $\mathbb{E}[\lambda\mu]$. Suppose first that $v\theta_R + (\gamma - \alpha)\mathbb{E}[\lambda\mu] \leq \overline{c} - \underline{c}$. Then radical protesters play $a_R = \mathbf{P}$ or $a_R = \mathbf{A}$, and a large protest $(\mathbf{P}, \mathbf{P})$ arises if and only if $\theta_M + \alpha\mathbb{E}[\mu] \geq \underline{c}$. If $v\theta_R + (\gamma - \alpha)\mathbb{E}[\lambda\mu] > \overline{c} - \underline{c}$, radicals play $a_R = \mathbf{V}$, and a large protest $(\mathbf{P}, \mathbf{V})$ arises if and only if $\theta_M + \alpha\mathbb{E}[\mu] - (\alpha + \beta)\mathbb{E}[\lambda\mu] \geq \underline{c}$. We therefore define

$$m = \begin{cases} \dfrac{\underline{c} - \theta_M}{\alpha} & \text{if } v\theta_R + (\gamma - \alpha)\mathbb{E}[\lambda\mu] \leq \overline{c} - \underline{c}, \\ \dfrac{\underline{c} + (\alpha + \beta)\mathbb{E}[\lambda\mu] - \theta_M}{\alpha} & \text{otherwise.} \end{cases}$$

This proves Part (i) of the lemma.

Fix $\mathbb{E}[\mu]$. If $\theta_M + \alpha\mathbb{E}[\mu] < \underline{c}$, a large protest never arises, so Part (ii) of the lemma is true when setting $r = -1$. If $\theta_M + \alpha\mathbb{E}[\mu] \geq \underline{c}$, then a large protest arises if and only if $v\theta_R + (\gamma - \alpha)\mathbb{E}[\lambda\mu] \leq \overline{c} - \underline{c}$ or $\theta_M + \alpha\mathbb{E}[\mu] - (\alpha + \beta)\mathbb{E}[\lambda\mu] \geq \underline{c}$. Letting

$$r = \max\left[\frac{\overline{c} - \underline{c} - v\theta_R}{\gamma - \alpha}, \frac{\theta_M + \alpha\mathbb{E}[\mu] - \underline{c}}{\alpha + \beta}\right]$$

proves Part (ii) of the lemma.

A.3 Proof of Proposition 1

Table A.1 describes all learning traps according to: (i) the equilibrium played in the long run; (ii) the equilibrium that would be played under complete information; (iii) the nature of the belief bias (relative to the true values $\tilde{\lambda}, \tilde{\mu}$) that sustains the incorrect equilibrium.

In the following we show how to obtain the conditions on long-run beliefs described in the last column of Table A.1. We fix a learning trap $(a, \chi, (\tilde{\lambda}, \tilde{\mu}))$, and distinguish cases as a function of the equilibrium a played.

First case: $a = (\mathbf{P}, \mathbf{V})$**.** The distribution $f(. \mid (\mathbf{P}, \mathbf{V}), \tilde{\lambda}, \tilde{\mu})$ identifies both $\tilde{\lambda}$ and $\tilde{\mu}$. The on-path consistency condition thus implies that $\chi = \delta_{\tilde{\lambda},\tilde{\mu}}$, and hence $a = a^*(\chi) = a^*(\delta_{\tilde{\lambda},\tilde{\mu}})$, which contradicts the fact that $(a, \chi, (\tilde{\lambda}, \tilde{\mu}))$ is a learning trap.

Second case: $a = (\mathbf{P}, \mathbf{P})$**.** The distribution $f(. \mid (\mathbf{P}, \mathbf{P}), \tilde{\lambda}, \tilde{\mu})$ identifies $\tilde{\mu}$, and thus $\mathbb{E}_\chi[\mu] = \tilde{\mu}$. The fact that $a = a^*(\chi) = (\mathbf{P}, \mathbf{P})$ then implies that $\theta_M + \alpha\tilde{\mu} \geq \underline{c}$, which in turn implies that $a^*(\delta_{\tilde{\lambda},\tilde{\mu}}) \in \{(\mathbf{A}, \mathbf{V}), (\mathbf{P}, \mathbf{V})\}$. That $a = a^*(\chi) = (\mathbf{P}, \mathbf{P})$ also implies

$$v\theta_R + (\gamma - \alpha)\mathbb{E}_\chi[\lambda]\tilde{\mu} \leq \overline{c} - \underline{c}. \tag{8}$$

In either case $a^*(\delta_{\tilde{\lambda},\tilde{\mu}}) = (\mathbf{A}, \mathbf{V})$ and $a^*(\delta_{\tilde{\lambda},\tilde{\mu}}) = (\mathbf{P}, \mathbf{V})$ we have

$$v\theta_R + (\gamma - \alpha)\tilde{\lambda}\tilde{\mu} > \overline{c} - \underline{c},$$

which together with Equation 8 implies $\mathbb{E}[\lambda] < \tilde{\lambda}$.

The learning trap can then be either $(\mathbf{A}, \mathbf{V})$ or $(\mathbf{P}, \mathbf{V})$ depending on whether $\theta_M + \alpha\tilde{\mu} - (\alpha + \beta)\mathbb{E}_\chi[\lambda]\tilde{\mu}$ is larger or smaller than $\underline{c}$. This captures Rows 5 and 8 of Table A.1.

Third case: $a = (\mathbf{A}, \mathbf{V})$**.** The distribution $f(. \mid (\mathbf{A}, \mathbf{V}), \tilde{\lambda}, \tilde{\mu})$ identifies $\tilde{\lambda}\tilde{\mu}$, and thus χ satisfies $\mathbb{E}_\chi[\lambda\mu] = \tilde{\lambda}\tilde{\mu}$. Since $a = a^*(\chi) = (\mathbf{A}, \mathbf{V})$, we have $v\theta_R + (\gamma - \alpha)\tilde{\lambda}\tilde{\mu} > \overline{c} - \underline{c}$. So, the only possible full-information equilibrium in a learning trap is $(\mathbf{P}, \mathbf{V})$, which arises if

$$\theta_M + \alpha\mathbb{E}_\chi[\mu] - (\alpha + \beta)\tilde{\lambda}\tilde{\mu} \leq \underline{c} < \theta_M + \alpha\tilde{\mu} - (\alpha + \beta)\tilde{\lambda}\tilde{\mu},$$

which implies $\mathbb{E}_\chi[\mu] < \tilde{\mu}$. This captures Row 4 of Table A.1.

Fourth case: $a = (\mathbf{A}, \mathbf{P})$**.** The distribution $f(. \mid (\mathbf{A}, \mathbf{P}), \tilde{\lambda}, \tilde{\mu})$ identifies $\tilde{\lambda}\tilde{\mu}$. Hence, $v\theta_R + (\gamma - \alpha)\tilde{\lambda}\tilde{\mu} \leq \overline{c} - \underline{c}$, and $\theta_R + \alpha\tilde{\lambda}\tilde{\mu} > \underline{c}$, which implies that the only possible learning trap is one where $(\mathbf{P}, \mathbf{P})$ should be played. This is the case if

$$\theta_M + \alpha \mathbb{E}_\chi[\mu] \leq \underline{c} < \theta_M + \alpha\tilde{\mu},$$

which implies $\mathbb{E}_\chi[\mu] < \tilde{\mu}$. This case covers Row 3 of Table A.1.

Fifth case: $a = (\mathbf{A}, \mathbf{A})$**.** No information is revealed, and inaction implies that $\theta_M + \alpha\mathbb{E}_\chi[\mu] \leq \underline{c}$ and $\theta_R + \alpha\mathbb{E}_\chi[\lambda\mu] \leq \underline{c}$.

If $(\mathbf{A}, \mathbf{P})$, $(\mathbf{A}, \mathbf{V})$ or $(\mathbf{P}, \mathbf{V})$ is the full-information equilibrium, then $\theta_R + \alpha\tilde{\lambda}\tilde{\mu} > \underline{c}$, which implies $\mathbb{E}_\chi[\lambda\mu] < \tilde{\lambda}\tilde{\mu}$. If $(\mathbf{P}, \mathbf{V})$ is the full-information equilibrium, we have in addition $\theta_M + \alpha\tilde{\mu} > \underline{c}$, which implies $\mathbb{E}_\chi[\mu] < \tilde{\mu}$. This captures Rows 1, 6 and 7 of Table A.1.

If $(\mathbf{P}, \mathbf{P})$ is the full-information equilibrium, then $\theta_M + \alpha\tilde{\mu} > \underline{c}$, which implies $\mathbb{E}_\chi[\mu] < \tilde{\mu}$. This captures Row 2 of Table A.1.

Table A.1: List of learning traps

#	Margin Affected	Self-Confirming Equilibrium	Full-Information Equilibrium	Long-Run Beliefs
1	Extensive	$(\mathbf{A}, \mathbf{A})$	$(\mathbf{A}, \mathbf{P})$	$\mathbb{E}_\chi[\lambda\mu] < \tilde{\lambda}\tilde{\mu}$
2		$(\mathbf{A}, \mathbf{A})$	$(\mathbf{P}, \mathbf{P})$	$\mathbb{E}_\chi[\mu] < \tilde{\mu}$
3		$(\mathbf{A}, \mathbf{P})$	$(\mathbf{P}, \mathbf{P})$	$\mathbb{E}_\chi[\lambda\mu] = \tilde{\lambda}\tilde{\mu}$, $\mathbb{E}_\chi[\mu] < \tilde{\mu}$
4		$(\mathbf{A}, \mathbf{V})$	$(\mathbf{P}, \mathbf{V})$	
5	Intensive	$(\mathbf{P}, \mathbf{P})$	$(\mathbf{P}, \mathbf{V})$	$\mathbb{E}_\chi[\mu] = \tilde{\mu}$, $\mathbb{E}_\chi[\lambda] < \tilde{\lambda}$
6	Both	$(\mathbf{A}, \mathbf{A})$	$(\mathbf{A}, \mathbf{V})$	$\mathbb{E}_\chi[\lambda\mu] < \tilde{\lambda}\tilde{\mu}$
7		$(\mathbf{A}, \mathbf{A})$	$(\mathbf{P}, \mathbf{V})$	$\mathbb{E}_\chi[\lambda\mu] < \tilde{\lambda}\tilde{\mu}$, $\mathbb{E}_\chi[\mu] < \tilde{\mu}$
8		$(\mathbf{P}, \mathbf{P})$	$(\mathbf{A}, \mathbf{V})$	$\mathbb{E}_\chi[\mu] = \tilde{\mu}$, $\mathbb{E}_\chi[\lambda] < \tilde{\lambda}$

A.4 Proof of Proposition 2

We start by characterizing $\Omega(\underline{c}, \overline{c})$ formally. Let $\Omega^{(\mathbf{P},\mathbf{P})}(\underline{c}, \overline{c})$ be the space of all pairs $[\chi, (\tilde{\lambda}, \tilde{\mu})]$ such that $[(\mathbf{P}, \mathbf{P}), \chi, (\tilde{\lambda}, \tilde{\mu})]$ is a self-confirming equilibrium, i.e. such that the

marginal of χ on μ is $\delta_{\tilde{\mu}}$ and

$$\begin{cases} \theta_M + \alpha\tilde{\mu} \geq \underline{c}, \\ v\theta_R + (\gamma - \alpha)\mathbb{E}_\chi[\lambda]\tilde{\mu} \leq \overline{c} - \underline{c}. \end{cases}$$

Similarly, let $\Omega^{(\mathbf{P},\mathbf{V})}(\underline{c},\overline{c})$ be the space of all pairs $[\chi,(\tilde{\lambda},\tilde{\mu})]$ such that $\chi = \delta_{\tilde{\lambda},\tilde{\mu}}$ and

$$\begin{cases} \theta_M + \alpha\tilde{\mu} - (\alpha+\beta)\tilde{\lambda}\tilde{\mu} \geq \underline{c}, \\ v\theta_R + (\gamma - \alpha)\tilde{\lambda}\tilde{\mu} > \overline{c} - \underline{c}. \end{cases}$$

We have $\Omega(\underline{c},\overline{c}) = \Omega^{(\mathbf{P},\mathbf{P})}(\underline{c},\overline{c}) \bigcup \Omega^{(\mathbf{P},\mathbf{V})}(\underline{c},\overline{c})$. Consider some variation in costs with $\underline{c}' < \underline{c}, \overline{c}' < \overline{c}$. We distinguish two cases.

Suppose first that $\overline{c}' - \underline{c}' \geq \overline{c} - \underline{c}$. Then take any $[\chi,(\tilde{\lambda},\tilde{\mu})] \in \Omega(\underline{c},\overline{c})$. We will show that $[\chi,(\tilde{\lambda},\tilde{\mu})] \in \Omega(\underline{c}',\overline{c}')$ as well. If $[\chi,(\tilde{\lambda},\tilde{\mu})] \in \Omega^{(\mathbf{P},\mathbf{P})}(\underline{c},\overline{c})$, then $[\chi,(\tilde{\lambda},\tilde{\mu})] \in \Omega^{(\mathbf{P},\mathbf{P})}(\underline{c}',\overline{c}')$, and thus $[\chi,(\tilde{\lambda},\tilde{\mu})] \in \Omega(\underline{c}',\overline{c}')$. The remaining case is the one where $[\chi,(\tilde{\lambda},\tilde{\mu})] \in \Omega^{(\mathbf{P},\mathbf{V})}(\underline{c},\overline{c})$. We then have

$$\theta_M + \alpha\tilde{\mu} - (\alpha+\beta)\tilde{\lambda}\tilde{\mu} \geq \underline{c} \Rightarrow \theta_M + \alpha\tilde{\mu} - (\alpha+\beta)\tilde{\lambda}\tilde{\mu} \geq \underline{c}' \text{ and } \theta_M + \alpha\tilde{\mu} \geq \underline{c}'.$$

Thus, $[\chi,(\tilde{\lambda},\tilde{\mu})] \in \Omega^{(\mathbf{P},\mathbf{V})}(\underline{c}',\overline{c}')$ if $v\theta_R + (\gamma-\alpha)\tilde{\lambda}\tilde{\mu} > \overline{c}' - \underline{c}'$, and $[\chi,(\tilde{\lambda},\tilde{\mu})] \in \Omega^{(\mathbf{P},\mathbf{P})}(\underline{c}',\overline{c}')$ otherwise (recall that $\mathbb{E}_\chi[\lambda] = \tilde{\lambda}$ since $[(\mathbf{P},\mathbf{V}),\chi,(\tilde{\lambda},\tilde{\mu})]$ is a self-confirming equilibrium). In both cases, we have $[\chi,(\tilde{\lambda},\tilde{\mu})] \in \Omega(\underline{c}',\overline{c}')$, which proves part (i) of the proposition.

Suppose now that $\overline{c}' - \underline{c}' < \overline{c} - \underline{c}$. Take any $[\chi,(\tilde{\lambda},\tilde{\mu})]$ such that $\chi = \delta_{\tilde{\lambda},\tilde{\mu}}$ and any parameters $(v,\theta_M,\theta_R,\alpha,\beta,\gamma)$ such that

$$\begin{cases} \theta_M + \alpha\tilde{\mu} \geq \underline{c}, \\ v\theta_R + (\gamma-\alpha)\tilde{\lambda}\tilde{\mu} \leq \overline{c} - \underline{c}, \\ \theta_M + \alpha\tilde{\mu} - (\alpha+\beta)\tilde{\lambda}\tilde{\mu} < \underline{c}', \\ v\theta_R + (\gamma-\alpha)\tilde{\lambda}\tilde{\mu} > \overline{c}' - \underline{c}'. \end{cases}$$

It is easy to check that this space is not empty. In addition, the first two inequalities imply that $[\chi,(\tilde{\lambda},\tilde{\mu})] \in \Omega^{(\mathbf{P},\mathbf{P})}(\underline{c},\overline{c})$ while the last two inequalities imply that $[\chi,(\tilde{\lambda},\tilde{\mu})] \notin \Omega(\underline{c}',\overline{c}')$. Hence, $\Omega(\underline{c}',\overline{c}') \not\supseteq \Omega(\underline{c},\overline{c})$. ∎

A.5 Proof of Proposition 3

Since $\theta_M + \alpha \mathbb{E}_{\chi_{2b}}[\mu] \geq \underline{c}$, the equilibrium can only be $(\mathbf{P}, \mathbf{P})$, $(\mathbf{A}, \mathbf{V})$ or $(\mathbf{P}, \mathbf{V})$. The equilibrium is: (i) $(\mathbf{P}, \mathbf{P})$ if and only if $v\theta_R + (\gamma - \alpha)\mathbb{E}_{\chi_{2b}}[\lambda\mu] \leq \overline{c} - \underline{c}$; (ii) $(\mathbf{P}, \mathbf{V})$ if and only if $v\theta_R + (\gamma - \alpha)\mathbb{E}_{\chi_{2b}}[\lambda\mu] > \overline{c} - \underline{c}$ and $\theta_M + \alpha \mathbb{E}_{\chi_{2b}}[\mu] - (\alpha + \beta)\mathbb{E}_{\chi_{2b}}[\lambda\mu] \geq \underline{c}$; and (iii) $(\mathbf{A}, \mathbf{V})$ otherwise. Letting

$$r_1 = \frac{\overline{c} - \underline{c} - v\theta_R}{\gamma - \alpha}$$

and

$$r_2 = \max\left[r_1, \frac{\theta_M + \alpha \mathbb{E}_{\chi_{2b}}[\mu] - \underline{c}}{\alpha + \beta}\right]$$

concludes. ∎

B Elements of Context

In 2015, President François Hollande decided to introduce a carbon tax on top of the existing gas tax to align the after-tax prices of diesel and gasoline. Despite rising oil prices since 2016 and increasing car-related expenses, the carbon tax was confirmed in 2017 by the newly elected President Emmanuel Macron. In January 2018, a few months later, Prime Minister Philippe lowered the speed limit on secondary roads from 90 km/h to 80 km/h, citing concerns about road safety. This decision, which was not included in Macron's campaign platform, led to numerous slowdowns across the country. The new 80 km/h regulation took effect on July 1st, 2018.

Despite growing discontent, especially among motorists, the annual increase in the carbon tax was confirmed in the 2019 budget at the end of the summer recess. In May 2018, a few months earlier, a motorist had started a petition against the gas tax on the Change.org platform. Though the petition received only a few hundred signatures during its first few months, it was mentioned in a local newspaper on October 12th, 2018. The newspaper had a local readership in Seine-et-Marne (a department on the outskirts of the Paris region), where the article triggered the first wave of signatures. The wife of a truck driver who was planning to block the Paris ring road in November saw the article and shared a link to the petition on Facebook. Nine days and thousands of local signatures later, a national newspaper published a new article about the petition and the planned roadblock, causing signatures to skyrocket nationwide. On October 24th, an online video recommended using yellow safety vests, which are required by law for all car owners to keep in their trunks, as a rallying symbol for angry drivers. The organizers of the roadblocks relied heavily on Facebook to spread the word, and several websites

were created to list relevant local Facebook groups. On November 17th, hundreds of thousands of protesters blocked hundreds of roads across France.

The movement resorted to more conventional weekly demonstrations in France's major cities, as most roadblocks were quickly removed. A peak of violence was reached on December 1st in Paris. The following Saturday, police tanks were mobilized and 2,000 people were arrested. On December 5th and 10th, as a sign of peace, President Macron announced that he would abandon the planned gas tax hike, then presented a 10 billion euro plan that significantly bent the government's budgetary policy. The main transfer to low-wage workers (*Prime d'Activité*) was both increased and expanded, which uniformly benefited all regions of France, independently of the extent of the mobilization (Leroy, 2024). He also called for the compilation of lists of grievances (*Cahiers de doléances*, as was done during the French Revolution in 1789) across the country, to be followed by hundreds of town hall meetings to allow everyone to voice their concerns through a "Great National Debate" (*Grand Débat National*).

Following this response, some roadblocks became permanent campsites, and weekly demonstrations continued for months. However, the number of protesters soon became negligible (except in Paris, where some large demonstrations still took place until March 2019, attracting protesters from other parts of France). At the same time, the protesters lost popular support and ultimately failed to present a united front for the upcoming elections (the 2019 European Parliament elections on May 26th). The movement remained active online in the following years, organizing sporadic protests where yellow vests were worn as a badge of honor. By 2024, it had become a trope to explain voting patterns, especially for far-right parties. As such, this simple piece of clothing has become an enduring and divisive icon in the French political landscape.

C Data Sources

C.1 Street Protests

A website (www.blocage17novembre.fr) was created to coordinate the mobilization. It provided a map of the organized blockades, updated in real-time. As of November 16, the map documented 788 geolocated blockades. We use this map to document the offline mobilization of the Yellow Vests, summarized in Panel C of Figure D.1. To check the validity of this source, we searched for all press reports of these events in the universe of newspaper articles published in November 18th and 19th (in national and regional daily newspapers) and we recover 613 articles. A less systematic search suggests that many of

the events we miss were still reported by local radio and TV channels or, later, by weekly local newspapers.

C.2 Change.org Petition

Change.org gave us access to an anonymized list of the signatories of the petition "Pour une baisse du prix des carburants à la pompe". Each observation is associated with the date of signature and the ZIP code of the signatory. We restrict the data to signatures in mainland France and with a valid ZIP code. By October 16, 2019, the petition had garnered 1,247,816 signatures, including 1,043,337 with a valid French ZIP code. We use the ZIP code to compute the signature rate in each municipality by dividing the number of signatures in each municipality by its population. When necessary, we allocate signatures associated to this ZIP code across relevant municipalities proportionally to population. Panel A in Figure D.1 shows the distribution of signature rates at the municipal level over France.

C.3 Facebook Activity

The main websites coordinating demonstrations listed local Facebook groups.[1] To document online mobilization, we looked for public Facebook groups and pages related to the movement. Due to the limitations of the Facebook API, we had to look for groups and pages manually, between December 12 and December 15, 2018 for groups and between March 21 and March 23, 2019 for pages. We used Netvizz to retrieve content between April 2 and April 10, 2019. Note that Netvizz did not allow us to retrieve actual discussions happening on Facebook groups. We use a keyword search approach to find Facebook groups and pages, performing requests on Facebook's search engine and manually retrieving results. These searches were performed using temporary sessions in order to minimize bias induced by Facebook's algorithm.

For groups, our aim was to retrieve as many groups linked to the Yellow Vests as possible. To this end, we started by searching for the keywords "gilet jaune" and "hausse carburant", on their own and associated with the the codes and names of the départements and of the former and current regions, as well as the names of all municipalities with more than 10,000 inhabitants.[2] Then, we performed further searches

[1] First blocage17novembre.fr, then gilets-jaunes.com and giletsjaunes-coordination.fr.

[2] Restricting the keywords used to these large municipalities is necessary as the number of municipalities in France is very high. It might introduce a bias towards groups associated to denser areas. Fortunately, this bias is reduced by a characteristic of Face-

with the keywords "hausse taxes", "blocage", "colere" and "17 novembre", associated with the names of the French départements, the names of the former and current regions, and the same list of municipalities as before. Finally, we performed searches for the following keywords: "gillet jaune", "gilets jaune", "manif 17 novembre", "manif 24 novembre", "manif 1 decembre", "manif 8 decembre", "macron 17 novembre", "macron 24 novembre", "macron 1 decembre", "macron 8 decembre","blocus 17 novembre", "blocus 24 novembre", "blocus 1 decembre", "blocus 8 decembre", "blocage 17 novembre", "blocage 24 novembre", "blocage 1 decembre", "blocage 8 decembre".[3]

For pages, as our aim was not to retrieve the universe of active Yellow Vests communities but simply a sample of messages large enough to perform text analysis, we relied on a smaller number of searches, searching for the keywords "gilet jaune" and "blocage hausse carburant" on their own or associated with the codes and names of the départements as well as a list of the largest cities.[4]

Yellow Vests Groups. For each group, we recorded the group's name, creation date, number of members, and number of publications. We eventually identified 3,033 groups in total, with over four million members. Over two-thirds of the groups were associated with a geographical area, and more than 40% of the total members belonged to these localized groups. Moreover, only 20% of the posts emanated from national groups, suggesting that localized groups were the most active type. Table C.1 presents descriptive statistics on the dataset. Panel B in Figure D.1 displays the spatial distribution of municipalities with at least one specific Facebook group.

Yellow Vests Pages. We identified 617 Facebook pages and used Netvizz to retrieve their content (Rieder, 2013): posts, comments, and interactions (such as likes and shares).[5] This corpus features over 121,000 posts, 2.1 million comments, and 21 million interactions. Since Netvizz did not provide user ids associated with scraped content, we scraped Facebook again in January 2022 and collected (de-identified) user ids. Approximately 30% of pages had been deleted by January 2022. On the remaining pages, we

book's algorithm: when searching for groups and pages associated with a municipality on the platform, Facebook also retrieves results associated to nearby municipalities.

[3]We reviewed all the search results manually to only keep the groups clearly associated with the mouvement.

[4]The complete list of further keywords used is the following: paris; marseille; lyon; toulouse; nice; nantes; strasbourg; montpellier; bordeaux; lille; rennes; reims; le havre; saint etienne; toulon; grenoble; dijon; angers; villeurbanne; le mans; nimes; aix en provence; brest; clermont ferrand; limoges; tours.

[5]Netvizz is no longer available since August 21st, 2019.

Table C.1: Characteristics of Facebook groups

Targeted Audience	Groups	Members	Publications
National	502 (63%)	2,372,217	255,131
Region	164 (81%)	244,930	135,857
County	717 (81%)	507,729	320,263
Municipality	1,638 (65%)	983,057	742,036
Total	3,033 (70%)	4,109,325	1,453,878

Notes: In the first column of this table, we show the number of Facebook groups for each geographic focus. We infer the group's targeted audience from its name. In parentheses, we indicate the share of the number of groups created after 11/17. Other columns show the total number of members and the total number of publications (this number is right-censored by Facebook at 10,000 publications per group). The last line ("Total") includes 12 "foreign" groups, 11 of which were created after 11/17, including 1,392 members and associated with 591 publications.

retrieved 46% of the original posts and 18% of the original comments for this second data retrieval (see Table C.2). We show in Figure E.5 that both datasets are quite similar in terms of predicted political affiliation and topics. They also display qualitatively similar trends, though the second dataset generally displays larger increases in radical attitudes (Figure E.6).

Table C.2: Comparison Between the Two Data Collections on Facebook Pages

Data Collection	Pages	Posts	Comments	Sentences	Users
First	617	120,242	1,936,921	2,860,427	—
Second	411	56,062	352,733	706,182	120,463

Notes: This table presents simple count metrics to compare the datasets resulting from our two data collections on Facebook pages.

C.4 Tweets of Politicians

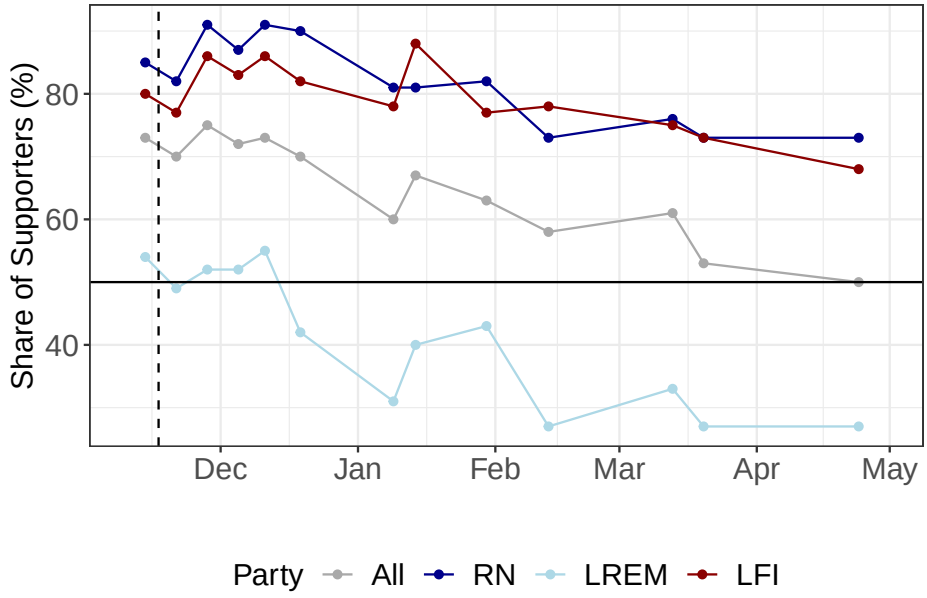
We built a dataset of tweets by politicians who belonged to the lower chamber of the French Parliament (the *Assemblée Nationale*) between 2017 and 2022. We consider the five largest French political parties: Rassemblement National (RN), Les Républicains (LR), La République en Marche (LREM), le Parti Socialiste (PS) and La France Insoumise (LFI). Politicians use Twitter to speak to their constituents directly. Thus, tweets are closer to daily social media messages than parliamentary speeches. They provide a natural, labeled dataset to train a machine learning classifier of party affiliation based on written

text. We then use our classifier to infer online protesters' political partisanship based on their Facebook messages. The complete list of politicians at the Assemblée Nationale is available on the official website of the Assemblée Nationale (see here). The dataset of French politicians on Twitter comes from the association "Regards Citoyens" (see here). We retrieved the last 3200 tweets of each politician via the Twitter API on December 12, 2021. The final dataset has 272 politicians for a total of 635,951 tweets.

C.5 Polls

The polling institute ELABE conducted several surveys between November 2018 and April 2019 for the news channel BFM TV. Figure C.1 reports their results on the evolution of popular support for the Yellow Vests movement.

Figure C.1: Evolution of the Popular Support for the Yellow Vests

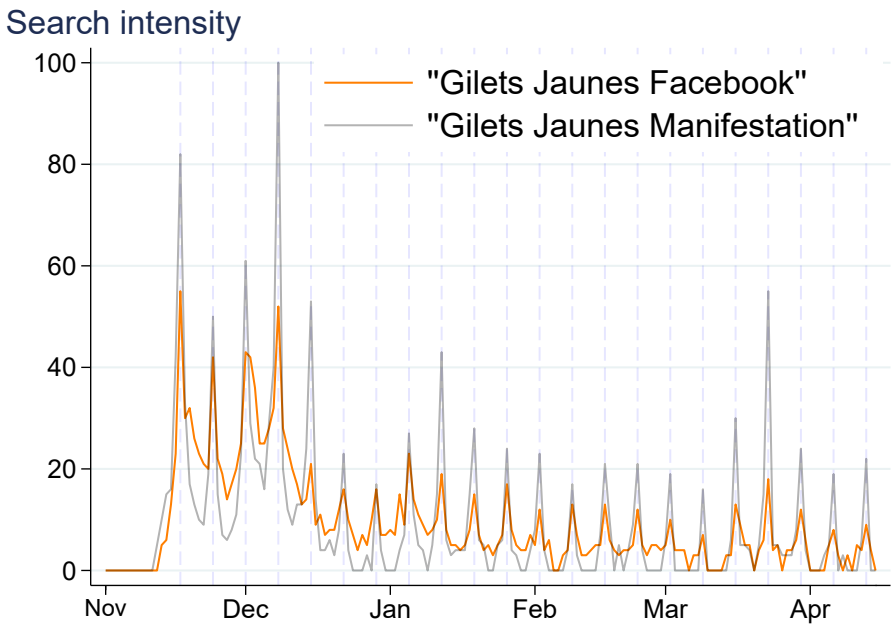


Notes: This figure plots the share of the population who declared they were supportive or sympathetic to the Yellow Vests movement over time. The vertical dashed line corresponds to 11/17. ELABE, the survey institute from which we collected data, conducted polls on 11/14/2018, 11/21/2018, 11/28/2018, 12/5/2018, 12/11/2018, 12/19/2018, 1/9/2019, 1/14/2019, 2/13/2019, 3/13/2019, 3/20/2019, and 4/24/2019. The number of respondents varies around 1,000 for the full sample and between 200 and 300 for the three subsamples, which correspond to declared vote during the first round of the 2017 presidential election. RN stands for "Rassemblement National" (far-right), LREM for President Macron's "La République En Marche" (center) and LFI for "La France Insoumise" (far-left).

C.6 Google Trends

Figure C.2 shows daily statistics from Google Trends in France for two phrases: *Gilets Jaunes Facebook* and *Gilets Jaunes Manifestation*. Street protests (*manifestation* in French) were organized every Saturday after 11/17. The weekly spikes in the second query may be driven by people trying to join the day's protest. However, a very similar pattern, both qualitatively and quantitatively, is observed for the first query, suggesting that the protests also triggered further attention to the Yellow Vest Facebook ecosystem. Before the first protest on 11/17, searches for *Gilets Jaunes Facebook* were virtually zero, even though some groups had been created for several weeks.

Figure C.2: Evolution of Google searches



Notes: Daily index of Google Search intensity in France for the keywords *Gilets jaunes Facebook* and *Gilets jaunes Manifestation* between November 1st, 2018 and April 15th, 2019. The dashed lines correspond to the weekly protests, starting in 11/17. *Source*: Google Trends.

D Supplement for the municipal analysis

D.1 Data at the municipal level

Some variables were only available at higher geographical levels. When relevant, we apportioned them according to municipal population.

Mobilization.

- **Blockade** is a dummy variable equal to 1 if there was a blockade in the municipality on 11/17.
- **Local group** is a dummy variable equal to 1 if a Facebook group was created in the municipality after 11/17.
- **Early signature** is the number of petition signatories per inhabitant on 11/16.
- **Early Groups** is the log number of local and regional groups (apportioned by municipal population) created before 11/17
- **Later Groups** is the log number of local and regional groups (apportioned by municipal population) created after 11/17
- **Members** is the log number of members who belong to later groups at the time of the scrape (mid-December 2018)
- **Posts** is the log number of messages posted on later groups at the time of the scrape (mid-December 2018)
- **Early Mobilization** is a dummy variable equal to 1 if the municipality belongs to the top quartile of municipalities in terms of Early signature and in terms of Early Groups.

Controls.

- **Baseline controls** includes the log population of the municipality and three binary variables equal to 1 if the municipality is an administrative center at the county (N=94), district (N=315) or subdistrict (N=1614) levels. *Source: Census (RP, complementary exploitation), 2016, INSEE*.

- **LZ** is a set of dummies for Living Zones. We cluster the 26 small municipalities that are alone in their Living Zone into a fictitious Living Zone in the estimation. *Source: INSEE.*

- **Population controls** includes two binary variables equal to 1 if the population of the municipality is larger than 20,000 or larger than 100,000 inhabitants, the share of immigrants in the population and the shares of the population in the following groups: 18-24 y.o.; 25-39 y.o.; 40-64 y.o.; over 65 y.o. *Source: Census 2016, INSEE.*

- **Geographical controls** includes the shares of the employed population commuting by car and public transportation, the median commuting distance, the share of roads where speed limit was lowered in 2018, as well as the share of diesel cars. *Source: Census 2016, INSEE. Déclarations Annuelles de Données Sociales (DADS), 2015, INSEE.*

- **Economic controls** includes the local unemployment rate, the fraction of employees with a non-permanent contract, log mean income, and the shares of the different *catégories socio-professionelles* defined by INSEE (executive, independent, middle-management, employee, manual worker) and the shares of the population without a high-school diploma, and with a university degree. *Source: Census 2016, INSEE. DADS, 2015, INSEE.*

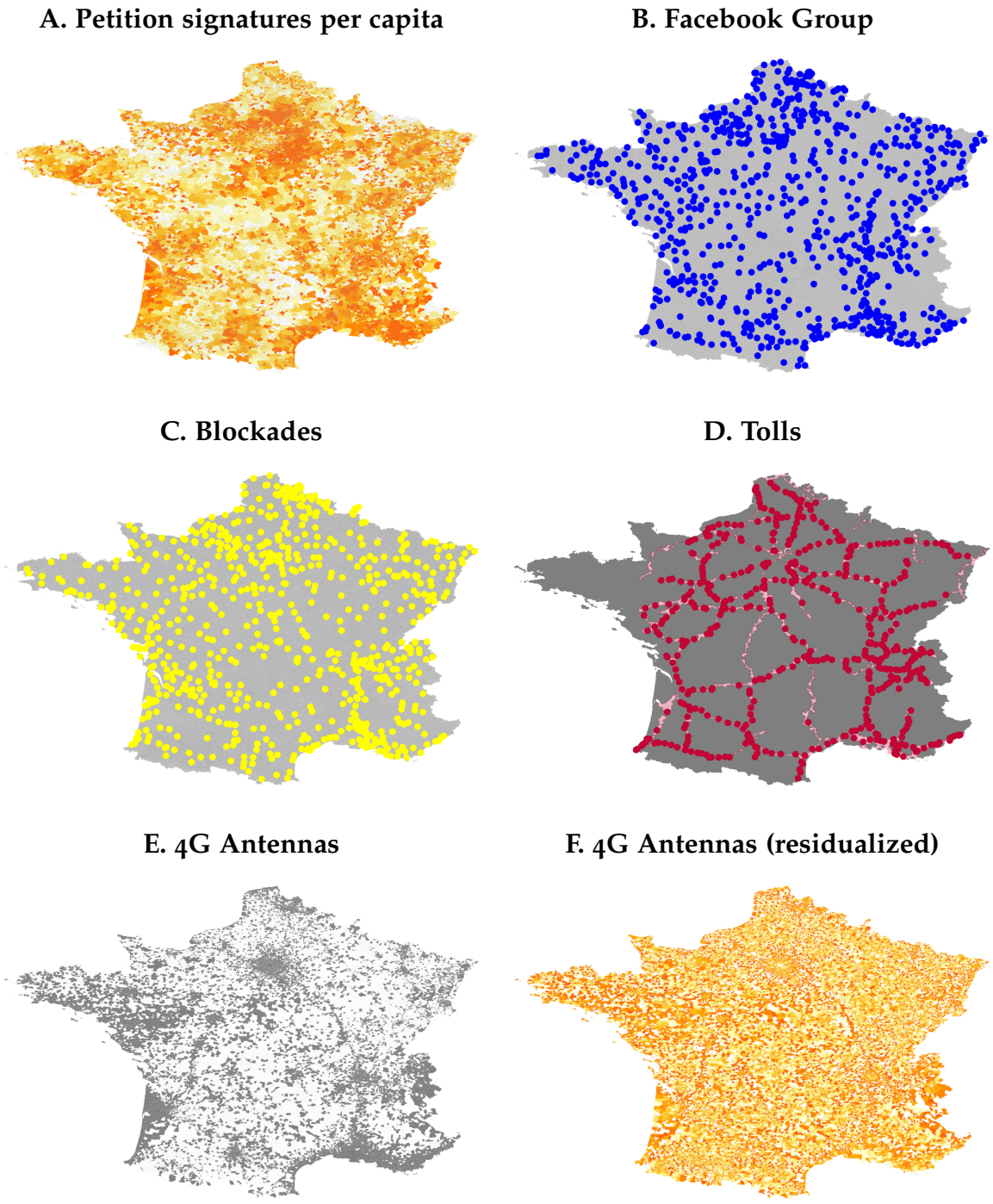
- **Political controls** includes the vote share for the five major candidates in the 2017 presidential election and the abstention rate. *Source: Ministry of the Interior.*

Instruments.

- **Tolls** is a dummy variable equal to 1 if the municipality hosts a highway toll in 2019. *Source: OpenStreetMap.*

- **4G Antenna** measures exposure to 4G as a dummy variable equal to 1 if the municipality has a working antenna prior to 11/17. *Source: Agence Nationale des Fréquences.*

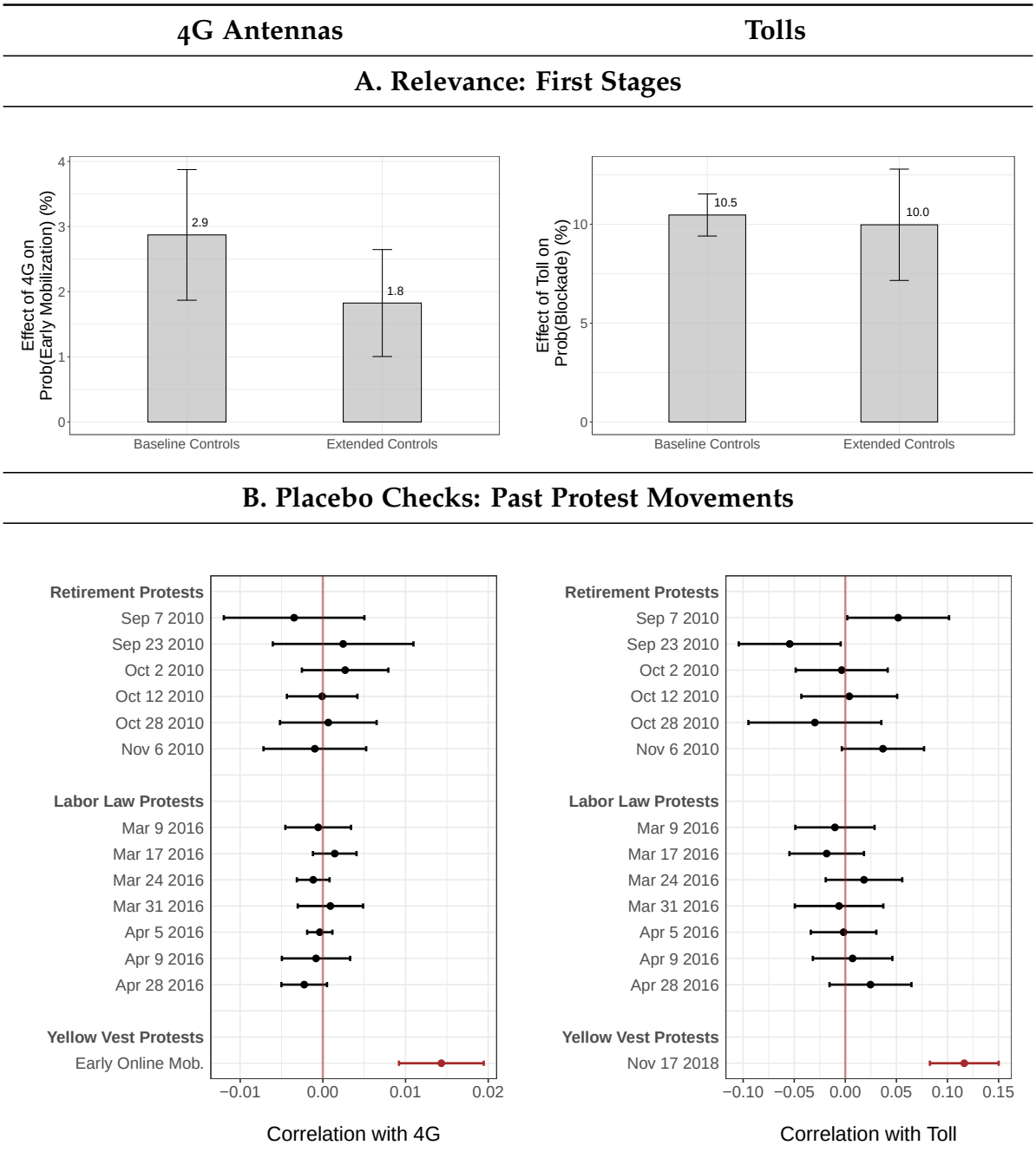
Figure D.1: Maps of Protests and Instruments



Notes: Panel A shows the distribution of the petition signature rate per inhabitant by the end of 2019. Panel B shows the location of municipalities with at least one municipal Yellow Vest Facebook group. Panel C shows the location of municipalities with at least one blockade on 11/17. Panel D shows the location of municipalities on the highway (in pink) and those with at least one highway toll (red dots). Panel E shows the location of municipalities with one 4G antenna before 11/17. Panel F shows the distribution of the previous binary variable after residualization on our set of extended controls.

D.2 Empirical strategy

Figure D.2: Relevance and Placebo Checks for the Instruments



Notes: Panel A shows the OLS coefficient estimate on the first stage regressions when we control for Baseline Controls and for a set of Extended Controls that comprise all variables listed as Controls in Appendix D.1. Panel B shows the OLS coefficient estimate and 95% confidence interval on the correlation between various binary variables if the municipality witnessed a protest in 2010 or 2016 on a given day, controlling for our set of Extended Controls. Standard errors are clustered at the LZ level.

D.3 Additional regression results

Table D.1: Feedback Loop — Robustness to alternative roadblock definition

	Dependent Variable:				
	Offline 11/17	Online Mobilization on Facebook Post-11/17			
	Blockade (indicator)	Group (indicator)	Groups (in logs)	Members (in logs)	Posts (in logs)
	(1)	(2)	(3)	(4)	(5)
Panel A: OLS					
Online Pre-11/17	0.020***				
	(0.005)				
Blockade		0.139***	0.346***	0.331***	0.333***
		(0.020)	(0.058)	(0.077)	(0.081)
Panel B: Reduced form					
4G Antenna	0.003**				
	(0.001)				
Toll		0.044***	0.110***	0.131***	0.125**
		(0.012)	(0.038)	(0.048)	(0.049)
Panel C: 2SLS					
Online Pre-11/17	0.173**				
	(0.087)				
Blockade		0.586***	1.472***	1.755***	1.672**
		(0.170)	(0.543)	(0.661)	(0.661)
Controls	✓	✓	✓	✓	✓
Mean dep. var.	0.02	0.02	-5.15	-0.01	-0.13
Observations	34475	34475	34475	34475	34475
Robust F Stat	19.09	34.27	34.27	34.27	34.27

Notes: Replication of Table 1 using a more restrictive definition of B_m whereby a municipality is considered as blocked if a protest was both announced on the eve of 11/17 and reported in a newspaper in the two following days. Panel A shows the OLS estimates of the correlation between O^e and B (Column 1) and between B and O^ℓ (Columns 2 to 5). Panel B shows the OLS estimates of the correlation between our outcome variables and our instruments: a binary variable equal to 1 if the municipality has a working 4G antenna before 11/17 and a binary variable equal to 1 if the municipality hosts a highway toll. Panel C shows the corresponding 2SLS estimates. Controls are listed in Appendix D.1 and include a set of over 1,600 Living Zone fixed effects. We cluster standard errors at the Living Zone level. *: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$.

Table D.2: 4G Instrument – Robustness

	Dependent Variable:						
	Blockade on 11/17 (indicator)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: OLS							
Online Pre-11/17	0.037***	0.033***	0.035***	0.035***	0.034***	0.034***	0.029***
	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.005)
Panel B: Reduced form							
4G Antenna	0.009***	0.006***	0.007***	0.007***	0.006***	0.006***	0.005***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Panel C: 2SLS							
Online Pre-11/17	0.265***	0.225***	0.273***	0.270***	0.258***	0.242***	0.233**
	(0.062)	(0.076)	(0.086)	(0.088)	(0.092)	(0.092)	(0.100)
Baseline Controls	✓	✓	✓	✓	✓	✓	✓
Living Zone Fixed Effect		✓	✓	✓	✓	✓	✓
Population Controls			✓	✓	✓	✓	✓
Geographical Controls				✓	✓	✓	✓
Economic Controls					✓	✓	✓
Political Controls						✓	✓
Excluding Paris Region							✓
Mean dep. var.	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Observations	34475	34475	34475	34475	34475	34475	33203
Robust F Stat	47.16	31.71	31.54	32.48	29.02	28.38	25.57

Notes: This Table shows specification and sample checks on the result shown in Column (1) in Table 1. Controls are listed in Appendix D.1. Column (6) corresponds to the full specification. In Column (7), we restrict the sample to municipalities outside the Paris region. We cluster standard errors at the Living Zone level. *: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$.

Table D.3: Toll Instrument – Robustness on the Likelihood of Creating a New Facebook Group

	Dependent Variable:								
	New Local Facebook Group (indicator)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: OLS									
Blockade	0.123***	0.122***	0.117***	0.116***	0.115***	0.114***	0.113***	0.107***	0.164***
	(0.015)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.017)	(0.032)
Panel B: Reduced form									
Toll	0.047***	0.044***	0.044***	0.044***	0.044***	0.044***	0.044***	0.041***	0.034***
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Panel C: 2SLS									
Blockade	0.446***	0.439***	0.442***	0.442***	0.443***	0.445***	0.445***	0.433***	0.413***
	(0.120)	(0.125)	(0.124)	(0.124)	(0.124)	(0.124)	(0.126)	(0.135)	(0.158)
Baseline Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
Living Zone Fixed Effect		✓	✓	✓	✓	✓	✓	✓	✓
Population Controls			✓	✓	✓	✓	✓	✓	✓
Geographical Controls				✓	✓	✓	✓	✓	✓
Economic Controls					✓	✓	✓	✓	✓
Political Controls						✓	✓	✓	✓
Online Pre-11/17							✓		
Excluding Paris Region								✓	
City Next to Highway									✓
Mean dependent variable	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.05
Observations	34475	34475	34475	34475	34475	34475	34475	33203	3905
Robust F Stat	49.21	47.28	47.12	48.02	48.29	48.34	48.05	42.69	31.22

Notes: This Table shows specification and sample checks on the result shown in Column (2) in Table 1. Controls are listed in Appendix D.1. Column (6) corresponds to the full specification. In Column (7), we control for the measure of early online mobilization (our main variable of interest in Equation (3)) and for the 4G Antenna dummy. In Column (8), we restrict the sample to municipalities outside the Paris region. In Column (9), we restrict the sample to municipalities that are located on a highway. We cluster standard errors at the Living Zone level. *: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$.

Table D.4: Toll Instrument – Robustness on the Number of New Facebook Groups

	Dependent Variable:								
	Number of New Facebook Groups (in logs)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: OLS									
Blockade	0.348***	0.309***	0.306***	0.302***	0.297***	0.295***	0.288***	0.273***	0.432***
	(0.055)	(0.046)	(0.047)	(0.046)	(0.046)	(0.046)	(0.046)	(0.049)	(0.092)
Panel B: Reduced form									
Toll	0.150***	0.110***	0.110***	0.111***	0.111***	0.113***	0.110***	0.106***	0.106***
	(0.049)	(0.038)	(0.038)	(0.038)	(0.038)	(0.038)	(0.038)	(0.039)	(0.041)
Panel C: 2SLS									
Blockade	1.431***	1.099***	1.108***	1.111***	1.112***	1.127***	1.117***	1.118**	1.272**
	(0.478)	(0.395)	(0.397)	(0.396)	(0.396)	(0.396)	(0.401)	(0.434)	(0.517)
Baseline Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
Living Zone Fixed Effect		✓	✓	✓	✓	✓	✓	✓	✓
Population Controls			✓	✓	✓	✓	✓	✓	✓
Geographical Controls				✓	✓	✓	✓	✓	✓
Economic Controls					✓	✓	✓	✓	✓
Political Controls						✓	✓	✓	✓
Online Pre-11/17							✓		
Excluding Paris Region								✓	
City Next to Highway									✓
Mean dependent variable	-5.15	-5.15	-5.15	-5.15	-5.15	-5.15	-5.15	-5.18	-4.24
Observations	34475	34475	34475	34475	34475	34475	34475	33203	3905
Robust F Stat	49.21	47.28	47.12	48.02	48.29	48.34	48.05	42.69	31.22

Notes: This Table shows specification and sample checks on the result shown in Column (3) in Table 1. Controls are listed in Appendix D.1. Column (6) corresponds to the full specification. In Column (7), we control for the measure of early online mobilization (our main variable of interest in Equation (3)) and for the 4G Antenna dummy. In Column (8), we restrict the sample to municipalities outside the Paris region. In Column (9), we restrict the sample to municipalities that are located on a highway. We cluster standard errors at the Living Zone level. *: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$.

Table D.5: Toll Instrument – Robustness on the Number of Members in New Facebook Groups

	Dependent Variable:								
	Number of Members in New Facebook Groups (in logs)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: OLS									
Blockade	0.396***	0.276***	0.279***	0.275***	0.271***	0.271***	0.264***	0.224***	0.428***
	(0.092)	(0.057)	(0.059)	(0.057)	(0.057)	(0.057)	(0.058)	(0.057)	(0.097)
Panel B: Reduced form									
Toll	0.225**	0.135***	0.135***	0.130***	0.130***	0.131***	0.128***	0.114**	0.138***
	(0.089)	(0.049)	(0.049)	(0.048)	(0.048)	(0.048)	(0.048)	(0.049)	(0.048)
Panel C: 2SLS									
Blockade	2.152**	1.357***	1.357***	1.307***	1.299***	1.312***	1.295***	1.203**	1.661***
	(0.850)	(0.499)	(0.496)	(0.490)	(0.490)	(0.490)	(0.496)	(0.527)	(0.616)
Baseline Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
Living Zone Fixed Effect		✓	✓	✓	✓	✓	✓	✓	✓
Population Controls			✓	✓	✓	✓	✓	✓	✓
Geographical Controls				✓	✓	✓	✓	✓	✓
Economic Controls					✓	✓	✓	✓	✓
Political Controls						✓	✓	✓	✓
Online Pre-11/17							✓		
Excluding Paris Region								✓	
City Next to Highway									✓
Mean dependent variable	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.04	1.02
Observations	34475	34475	34475	34475	34475	34475	34475	33203	3905
Robust F Stat	49.21	47.28	47.12	48.02	48.29	48.34	48.05	42.69	31.22

Notes: This Table shows specification and sample checks on the result shown in Column (4) in Table 1. Controls are listed in Appendix D.1. Column (6) corresponds to the full specification. In Column (7), we control for the measure of early online mobilization (our main variable of interest in Equation (3)) and for the 4G Antenna dummy. In Column (8), we restrict the sample to municipalities outside the Paris region. In Column (9), we restrict the sample to municipalities that are located on a highway. We cluster standard errors at the Living Zone level. *: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$.

Table D.6: Toll Instrument – Robustness on the Number of Messages Posted on New Facebook Groups

	Dependent Variable:								
	Number of Messages Posted on New Facebook Groups (in logs)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: OLS									
Blockade	0.337***	0.268***	0.273***	0.269***	0.265***	0.265***	0.259***	0.214***	0.407***
	(0.085)	(0.060)	(0.061)	(0.059)	(0.060)	(0.059)	(0.060)	(0.059)	(0.097)
Panel B: Reduced form									
Toll	0.148*	0.132***	0.132***	0.126***	0.125**	0.126***	0.122**	0.106**	0.132***
	(0.084)	(0.049)	(0.049)	(0.049)	(0.049)	(0.049)	(0.049)	(0.049)	(0.049)
Panel C: 2SLS									
Blockade	1.411*	1.329***	1.321***	1.262**	1.250**	1.261**	1.239**	1.123**	1.589**
	(0.785)	(0.502)	(0.500)	(0.493)	(0.493)	(0.492)	(0.498)	(0.523)	(0.624)
Baseline Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
Living Zone Fixed Effect		✓	✓	✓	✓	✓	✓	✓	✓
Population Controls			✓	✓	✓	✓	✓	✓	✓
Geographical Controls				✓	✓	✓	✓	✓	✓
Economic Controls					✓	✓	✓	✓	✓
Political Controls						✓	✓	✓	✓
Online Pre-11/17							✓		
Excluding Paris Region								✓	
City Next to Highway									✓
Mean dependent variable	-0.13	-0.13	-0.13	-0.13	-0.13	-0.13	-0.13	-0.17	0.87
Observations	34475	34475	34475	34475	34475	34475	34475	33203	3905
Robust F Stat	49.21	47.28	47.12	48.02	48.29	48.34	48.05	42.69	31.22

Notes: This Table shows specification and sample checks on the result shown in Column (5) in Table 1. Controls are listed in Appendix D.1. Column (6) corresponds to the full specification. In Column (7), we control for the measure of early online mobilization (our main variable of interest in Equation (3)) and for the 4G Antenna dummy. In Column (8), we restrict the sample to municipalities outside the Paris region. In Column (9), we restrict the sample to municipalities that are located on a highway. We cluster standard errors at the Living Zone level. *: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$.

Table D.7: 4G Instrument – Robustness on the definition of early online mobilization

	Dependent Variable:								
	Blockade on 11/17 (indicator)								
Quantile cutoff	1/2	2/3	3/4	4/5	5/6	6/7	7/8	8/9	9/10
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: OLS									
Online Pre-11/17	0.006	0.016***	0.034***	0.049***	0.061***	0.065***	0.076***	0.072***	0.084***
	(0.004)	(0.005)	(0.006)	(0.009)	(0.011)	(0.012)	(0.013)	(0.013)	(0.016)
Panel B: 2SLS									
Online Pre-11/17	-1.139	0.241**	0.242***	0.302***	0.367***	0.433***	0.491***	0.690**	0.842**
	(1.419)	(0.096)	(0.092)	(0.114)	(0.140)	(0.167)	(0.189)	(0.288)	(0.361)
Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mean indep. var.	0.27	0.13	0.08	0.05	0.04	0.03	0.02	0.02	0.01
Observations	34475	34475	34475	34475	34475	34475	34475	34475	34475
Robust F Stat	0.70	21.25	28.38	25.78	22.66	22.28	21.53	13.72	12.31

Notes: This Table shows specification checks on the result shown in Column (1) in Table 1. Controls are listed in Appendix D.1. In each column, the definition of the independent variable changes: it is a binary variable equal to 1 if both the number of local groups and the signature rate before 11/17 are in the top half of municipalities (Column 1), in the top third (Column 2), up to the top decile (Column 10). Column (3) corresponds to our baseline definition. We cluster standard errors at the Living Zone level. *: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$.

Table D.8: Toll instrument – Robustness to using two instruments

	Dependent Variable:							
	Group (indicator)		Groups (in log)		Members (in log)		Posts (in log)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: 2SLS								
Blockade	0.446***	0.445***	0.952**	1.064***	0.688	1.087**	0.875*	1.122**
	(0.127)	(0.114)	(0.406)	(0.365)	(0.462)	(0.434)	(0.479)	(0.443)
Panel B: First-stage								
No other toll in LZ	0.133***	0.063**	0.133***	0.063**	0.133***	0.063**	0.133***	0.063**
	(0.022)	(0.029)	(0.022)	(0.029)	(0.022)	(0.029)	(0.022)	(0.029)
Toll		0.070***		0.070***		0.070***		0.070***
		(0.018)		(0.018)		(0.018)		(0.018)
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Mean dependent variable	0.02	0.02	-5.15	-5.15	-0.01	-0.01	-0.13	-0.13
Observations	34475	34475	34475	34475	34475	34475	34475	34475
Robust F Stat	35.32	25.24	35.32	25.24	35.32	25.24	35.32	25.24
P-value Hansen	.	0.99	.	0.60	.	0.14	.	0.36

Notes: The outcome and explanatory variables are described in the text. Panel A shows the 2SLS estimates and Panel B shows the first-stage OLS estimates. Controls are the same as in Table 1. "No other toll in LZ" is a binary variable equal to 1 if there is no toll in other municipalities of the Living Zone. We cluster standard errors at the Living Zone level. *: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$.

E Supplement for the analysis on Facebook pages

E.1 Text Pre-processing

We process all text corpora in the same way. We remove emojis, links, accents, punctuation, social media notifications (e.g., "Yellow Vests changed their profile picture"), and stopwords from the corpus. We also lowercase the text and lemmatize words. We keep hashtags, user mentions, verbs, nouns, proper nouns, adjectives, and numbers. We drop all tokens that occur less than ten times in the Facebook corpus.[6] This leaves us with approximately 40,000 unique tokens in the corpus. Most documents in our corpora are short text snippets (e.g., a phrase or a sentence). Some are longer and span over multiple sentences (e.g., Facebook posts). To keep all documents comparable, we work with unigrams at the sentence level.

E.2 Topic Model

The standard approach for topic modeling in the text as data literature is to rely on Latent Dirichlet Allocation (LDA) or one of its variants. LDA models documents as a distribution over multiple topics. Though this is often a reasonable assumption, it is implausible in the case of short text snippets (such as sentences), which often refer to only one topic (Yan, Guo, Lan and Cheng, 2013). For this reason, standard topic models are known to perform poorly on such short texts. As an alternative, we build a custom topic model in the spirit of Demszky et al. (2019). First, we produce word embeddings for the corpus and represent each sentence as a vector in the embedding space. We train a Word2Vec model using Gensim's implementation, with moving windows of eight tokens and ten iterations of training. We build sentence embeddings as the weighted average of the constituent word vectors, where the weights are smoothed inverse term frequencies (to assign higher weights to rare/distinctive words) (Arora, Liang and Ma, 2017). The resulting embedding space allows for a low-dimensional representation of text in which phrases that appear in similar contexts are located close to one another. Second, we group sentence vectors together into a small set of clusters. The goal is to have different clusters for different topics in the text. We rely on the K-Means algorithm. We train the algorithm on 100,000 randomly drawn sentences and predict clusters for the rest of the

[6]The frequency threshold does not influence results, but allows us to remove many uncommon spelling mistakes and other idiosyncrasies related to social media data.

corpus. We use the ten closest words to the cluster centroids to manually label topics.[7]

To further inspect the results of the topic model, Table E.1 shows the closest phrase to the centroid of each topic below. These phrases may be understood as the most representative text snippet for each topic. Similarly, Figure E.1 shows wordclouds for each topic. We choose to work with 15 topics for our main results. However, since the number of topics is a hyperparameter in our topic model, we also present resulting topics when specifying 5, 10, and 20 clusters (see Table E.2).

E.3 Sentiment Analysis

To measure emotional content in Facebook messages, we use a dictionary-based approach that assigns to a sentence a sentiment score ranging from -1 (very negative) to 1 (very positive). For each sentence, the sentiment score is obtained as the average of the sentiment scores of its constituent words. We rely on the VADER (Valence Aware Dictionary for Sentiment Reasoning) library for our main results. Table E.4 shows five of the most negative and five of the most positive sentences according to the VADER sentiment analyzer.

Our measure of sentiment could vary depending on the dictionary used. As a robustness check, we rely on French TextBlob as an alternative dictionary for word sentiment. We find that the VADER dictionary's density has larger tails as it tends to classify more sentences to the extremes of the sentiment spectrum. Nonetheless, both measures suggest an increase in average negative sentiment between November 2018 and March 2019. Figure E.3 decomposes the increase in average negative sentiment (as measured by TextBlob) using the method outlined in Section 3.3. Results are qualitatively similar to the main text results.

Robustness: emoticons. The classical approach to sentiment analysis has some drawbacks in our context. First, irony (a well-known feature of the French psyche) can lead to poor predictions. The following messages may be classified as positive by the method described above despite being negative: "Making America Great Again gave us everything but good"; "Congratulations to the government, #1 in keeping peaceful demonstrators out of the streets". Second, training sets in French are not as widely available as in English, and they are often extracted from very different contexts (for example, movie reviews).

[7] We also considered alternative labeling options, such as term frequency-inverse cluster frequency, which yield similar results.

Table E.1: Results of the Topic Model: Most representative phrases

Topic Human Label	Most representative phrase
Critiques	*visiblement representer peuple francais devenir lamentable attitude mepris*
Insults	*sale batard hont francais macron bouffon macron batard degage fumier*
Diffusion	*vouloir publier information verifier site diffuser savoir etre derriere info*
Towns-Hours	*samedi 5 janvier rdv 10h place verdun marche rdv 18h zenith pau partir convoi tarbes depart 18h30 max 19h co voiturage voir place*
Conspiracy	*souverainiste racisme fascisme etre frontal pensee correct tourner nation occidental homme blanc judeo chretien etre utilise arme psychologique mediatique tres puissant hegemonie moral ideologique pouvoir perdurer peuple europeen culpabiliser gauche systematiquement instrumentaliser ad horreur second guerre mondial discrediter national lui meme homme blanc nom jamais devoye*
Concerns	*2000 euro concerne restaurer service public disparu poste hopital maternite ecole instauration revenu minimum lieu aide diffus demander complexe limitation salaire 10 smic augmentation salaire meme proportion gros salaire reprise dette banque france banque prive limitation montant demander maison retraite ecole vraiment gratuite fourniture activite livre gratuit lieu donner aide servir chose detail complet utilisation impot blocage tipp salaire elu 4 smic fin privilege egalite transparence fonds*
Actions	*malheureusement laisse choix vouloir change aller falloir arreter pacifiste attendre roi rigoler voir faire defoncer tomber nuit*
Foreign Languages	*marie jo laziah*
Names	*rajoute prenom chaine rose annick patricia nelly angel sophia mary didier gabrielle maya pierre fanny magali ludivine isabelle nicole nathan marie patricia jeannine serge josiane eric marie fleur rose laly severine emilie delphine nanou ophelie yohann laurer nanou aya magdalena aurelie angele chantal fanny carine brigitte yael sylvie virginie dominique rachel frederic audrey benjamin marie jeanne phil laurence rachel jeremy annie patricia agnes nini*
Violence	*france ordre pouvoir continuer agresser impunite civil etre legitime defense cas attaque voir rue tv journaliste faire photo etre blesser flashball coup venir porter plainte ordre justement*
Other	*oui faire accord jean michel*
Politics	*faire site internet permettre inscrire revendication monde pouvoir proposer soutenir d lier etre veritable logique fin possibilite revendiquer systeme constitution battre revolte revolutionnaire systeme place deja logique pre institution etre legitimer adhesion populaire*
Support	*bonjour lilly cur courage etre fille formidable faire gros bisou*
Places	*79 44 85 16 13 80 06 01 53 36 69 bcp 17*
Food-Objects	*jamais faire greve vie etre fan kro merguez pis odeur pouilleux sentir pisse odeur pneu cramer*

Notes: For each topic, we present the closest phrase to the cluster centroid as measured by cosine similarity. We present the pre-processed (as opposed to raw) phrases.

Figure E.1: Results of the Topic Model: Wordclouds

A. Places [2.9%] B. Towns and hours [4.4%] C. Support [3.9%]

D. Diffusion [4.8%] E. Economic concerns [5.9%] F. Political institutions [7.6%]

G. Food and objects [6%] H. Critiques [6%] I. Insults [4.5%]

J. Violence [5.9%] K. Conspiracy [5.9%] L. Actions [7%]

M. Names [6.6%] N. Foreign Languages [8.4%] O. Other [20%]

Notes: This figure shows wordclouds associated with the fifteen topics we identify in our corpus. The size of words is determined by a term frequency-inverse document frequency (TF-IDF) metric, where each document is the entire collection of sentences associated with a given topic. This metric gives higher scores to words that are (i) more frequent in the corpus and (ii) particularly meaningful for each topic. Wordclouds are boxed inside a rectangle when the average sentiment of messages in the topic is negative. Squared brackets indicate the topic frequency (computed as the share of total messages in the corpus).

Table E.2: Results of the Topic Model for Alternative Numbers of Clusters

Panel A: Results of the Topic Model for 5 clusters

Topic	Most representative words
1	04, nimes, arras, nime, 77, narbonne, albi, chambery, 47, orleans
2	pouvoir, etre, consequent, favoriser, necessaire, n, global, politique, specifique, constitue
3	merde, connard, salopard, pourriture, encule, putain, hont, honte, batard, ordure
4	gabin, live, sympa, app, brancher, stp, ramous, cool, stabilisateur, coupure
5	laziah, misfortune, #noussommesgiletsjaune, dellacherie, exhort, substituons, sansone, pajalo, victory, naeim

Panel B: Results of the Topic Model for 10 clusters

Topic	Most representative words
1	etre, n, peuple, meme, politique, faiblesse, nefaste, veritable, gouvernement, destructeur
2	annuel, beneficiaire, compenser, bonus, salaire, taxation, production, exoneration, delocalisation, embauche
3	cr, flic, flics, policier, gazer, projectile, charger, manifestant, matraque, gendarme
4	zappe, zapper, tpmp, humoriste, fakenew, interviewe, conversation, cnew, interviewer, bfmtv
5	orlane, magdalena, grilo, correia, gourdon, leal, caudrelier, malaury, macedo, khaye
6	connard, merde, encule, bouffon, conard, pd, salope, enculer, fdp, batard
7	adhesion, charte, valider, definir, modalite, eventuel, prealable, specifique, necessaire, proposer
8	04, nimes, arras, albi, nime, royan, 77, narbonne, chambery, 47
9	courage, courag, bravo, felicitation, formidable, bisou, bisous, genial, soutien, continuation
10	sansone, dutie, facilitate, soldiers, auv, weier, unterstutzen, #jiletsjaune, ausbeutung, seem

Panel C: Results of the Topic Model for 20 clusters

Topic	Most representative words
1	beneficiaire, compenser, salaire, bonus, annuel, exoneration, plafonner, taxation, embauche, reduction
2	omo, #noussommesgiletsjaune, laziah, houpette, nooooon, jeoffrey, chab, limitatif, exhort, cageot
3	aller, faire, voir, la, etre, oui, vraiment, merde, savoir, meme
4	englos, royan, sisteron, pontivy, arras, seclin, hendaye, douai, roanne, albi
5	twitter, diffuse, info, publier, fb, diffuser, relater, page, interview, information
6	adhesion, structuration, proposer, proposition, definir, charte, structurer, concertation, revendication, necessaire
7	maud, johanna, gomes, anai, melanie, gregory, rudy, armand, melissa, mathias
8	bisous, courage, felicitation, courag, bisou, bravo, formidable, soutien, genial, coucou
9	asservissement, domination, peuple, deposseder, destructeur, gouvernance, oppression, politique, veritable
10	recours, illegal, sanction, infraction, poursuite, condamnation, delit, penal, abusif, commettre
11	41, 52, 58, 47, 38, 61, 69, 37, 46, 82
12	canette, chaussette, bouteille, cendrier, plastique, peintur, toilette, saucisson, scotch, brosse
13	cr, flic, flics, frapper, tabasser, matraquer, policier, gazer, matraque, tabasse
14	mafieux, imposteur, larbin, escroc, acolyte, magouilleur, maffieux, corrompu, dictateur, sbire
15	kassav, akiyo, diritti, sempr, dittaturer, etait, popolo, quando, anch, infami
16	stupide, pathetique, affliger, pitoyable, malsain, stupidite, abject, irrespectueux, insultant, grossier
17	15h, 17h30, 16h30, 10h, 14h00, 11h, gare, 8h30, 18h, 18h30
18	laziah, #noussommesgiletsjaune, gourdon, misfortune, orlane, grilo, victory, duquesnoy, dellacherie, macedo
19	#jiletsjaune, created, soldiers, #assembleenationale, #coletesamarelo, #parisprotest, dutie, unterstutzen, #france3
20	connard, encule, batard, salope, fdp, merde, conard, enculer, pd, salopard

Notes: This table presents the top words associated with our topics when requesting alternative numbers of clusters (respectively 5, 10, and 20). For each topic, we report the closest words to the cluster centroid (measured by cosine similarity).

Table E.3: Examples of Pro-violence and Anti-violence Phrases

Panel A: Online protester phrases in favor of violence

C'est la violence des casseurs et les degats qu'ils ont fait qui font plier, un peu, Macron... et malheureusement pas nos manifs. It's the violence of the rioters and the damage they've done that's making Macron bend a little... and unfortunately not our demonstrations!

c'est vraiment honteux de nous sortir de telles mesures maintenant, ils restent sourds et poussent a la violence. it's really shameful to come up with such measures now, they remain deaf and push for violence.

Et meme si certains vous taxent d'etre des violents, continuez, la violence, c'est comme la chimiotherapie, personne ne la fait de gaiete de coeur, ce n'est pas un amusement, mais c'est une epreuve. And even if some criticize you for being violent, keep it up, violence is like chemotherapy, no one does it gladly, it's not fun, but it's a trial.

Nous ca fait depuis le 17 novembre, il y a de la casse et de la violence et on a rien obtenu car on est pas assez nombreux. Since November 17, there's been breakage and violence, and we've achieved nothing because there aren't enough of us.

Pacifistes et utopistes vous ne servez a rien! Restez chez vous ou vous vous ferez matraquer comme nous et pour rien par ces chiens que sont ces policiers qui continuent a servir l etat au detriment de leurs propres droits et des notres! Vous n etes pas dans la realite de notre pays. Aujourd hui encore nous sommes oblige de ressortir et de faire appel a nos traditions de violence pour defendre notre droit a une vie decente Pacifists and utopians, you're useless! Stay at home or you'll be bludgeoned like the rest of us and for nothing by those police dogs who continue to serve the state to the detriment of their own rights and ours! You're out of touch with the reality of our country. Even today, we are obliged to call on our traditions of violence to defend our right to a decent life.

Panel B: Online protester phrases opposed to violence

Il faudrait aussi peut-etre condamner les violences car c'est un reproche qui est fait perpetuellement aux gilets jaunes. Perhaps we should also condemn violence, as this is a criticism that is perpetually levelled at the Yellow Vests.

je vous soutiens et suis entierement d accord avec vous sauf sur la violence de ce week end mais tout le monde le deplore. I support you and agree with you wholeheartedly, except for this weekend's violence, which everyone deplores.

Des gens s'etonnent de constater la remontee d'Emmanuel Macron dans les sondages... Pouvions nous valablement penser que le soutien populaire du debut durerait eternellement dans le contexte actuel ? Je veux dire dans un contexte ou la violence recurrente People are surprised to see Emmanuel Macron's rise in the polls... Could we reasonably think that the initial popular support would last forever in the current context? I mean, in a context of recurring violence

G ete manifester pour la 1ere fois a bdx avec les gilets jaunes. Je suis arrivee un peu anxieuse et desespere et peur de la violences des debordements par la Situation de notre pays. I went to protest for the first time in Bordeaux with the Yellow Vests. I arrived a little anxious and despairing and afraid of the violence of the excesses by the situation of our country.

je ne suis pas pour la violence parceque c'est ce qui sabote le mouvement I'm not for violence because that's what sabotages the movement.

Soutien au peuple soyez prudents pas de violence SVP Support the people be careful no violence please

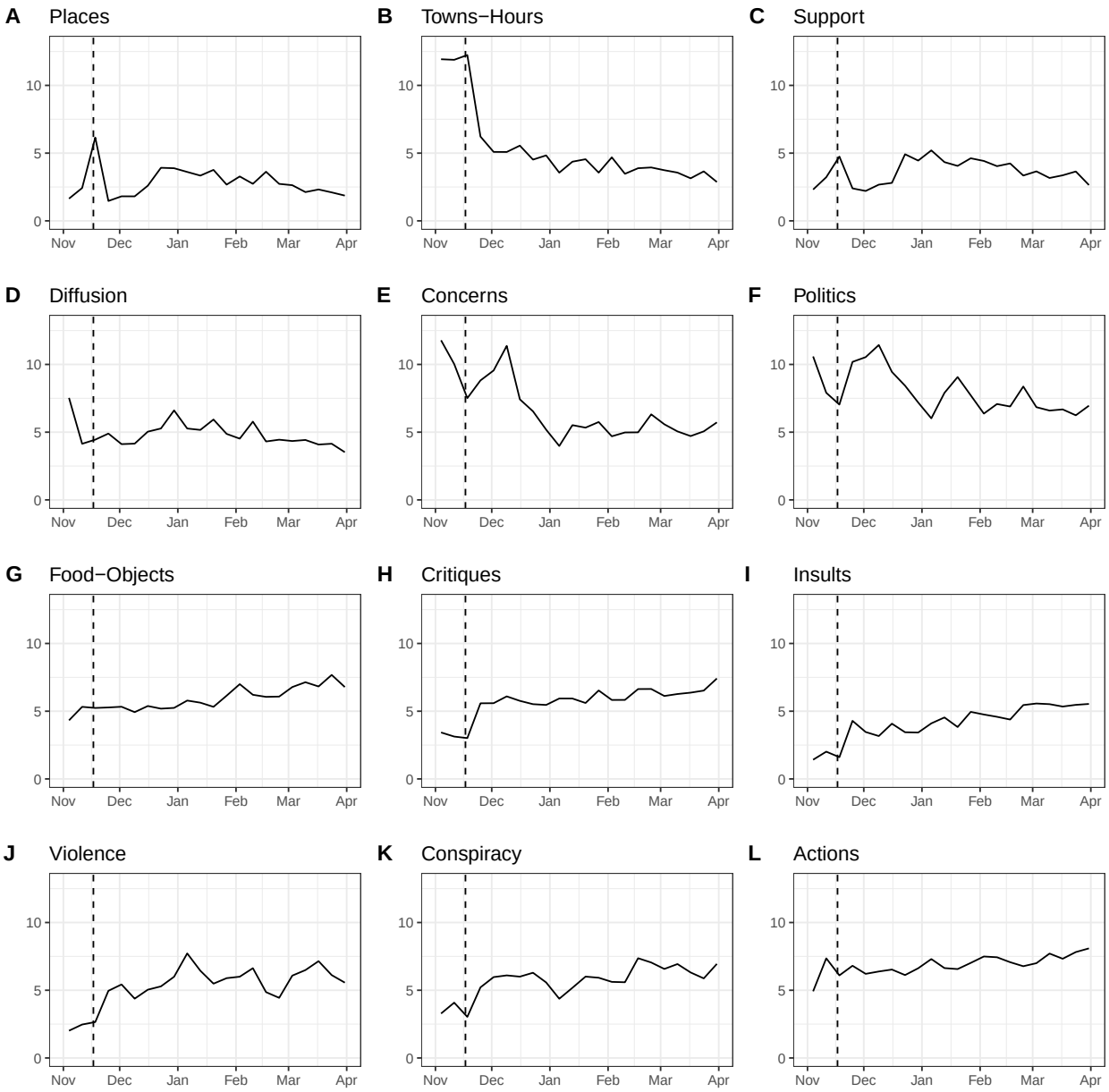
Il faut arreter de prendre des gants avec cette violence et la denoncer franchement. We have to stop taking the gloves off with this violence and denounce it frankly.

C'est horrible . Apres je sais pas ce qu'ils ont fait pour en arriver a ca mais la violence c'est jamais la bonne solution. It's horrible. I don't know what they did to get there, but violence is never the right solution.

Je ne soutien pas la violence, etant non violent moi meme. I don't support violence, being non-violent myself.

Notes: Selection of raw phrases that contain the token "violence". The original phrases in French are in italics. Their English translation follows.

Figure E.2: Topic Shares in Facebook Discussions Over Time



Notes: This figure shows weekly shares of the twelve topics of interest shown in Figure E.1. For all topics, the vertical dashed line corresponds to 11/17. The share of messages associated with violence is below 2.5% in early November and is consistently above 5% after December 10.

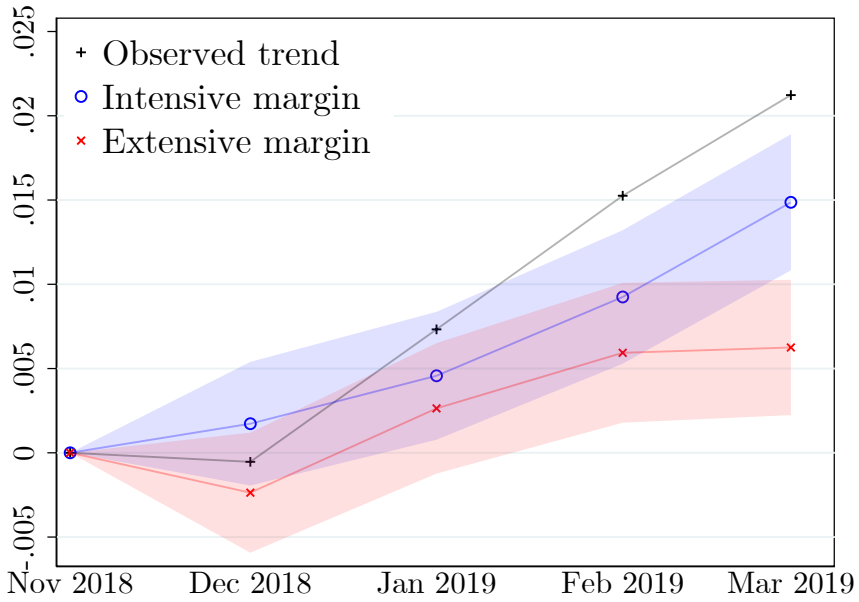
To overcome these problems, we take advantage of the fact that users can react to Facebook posts, using the following reactions: *love, haha, wow, angry, sad*. For each post in our corpus, we compute the weekly share of each of these reactions, displayed in Figure E.4. The share of *angry* reactions goes from 20% to almost 50% in less than three weeks, and remains stable in the following months.

Table E.4: Examples of positive and negative sentences

Sentiment	Sentence
Positive	*honneur gilet jaune* honor yellow vest
	mdr lol
	bravo congrats
	mercii jeune meilleur facon aider progres meilleur monde
	thanks young best way to help progress better world
	bravo gabin media honnete souhaite reussite merite equipe bravo gj
	congrats gabin honest media wish you success deserve team congrats yellow vest
Negative	*macron demission* macron resignation
	macron cabanon castananer enfer macron prison castaner hell
	florence menteur florence liar
	bande pourriture batard group of **** ****
	castaner assassin degage voleur menteur
	castaner murderer get out thief liar

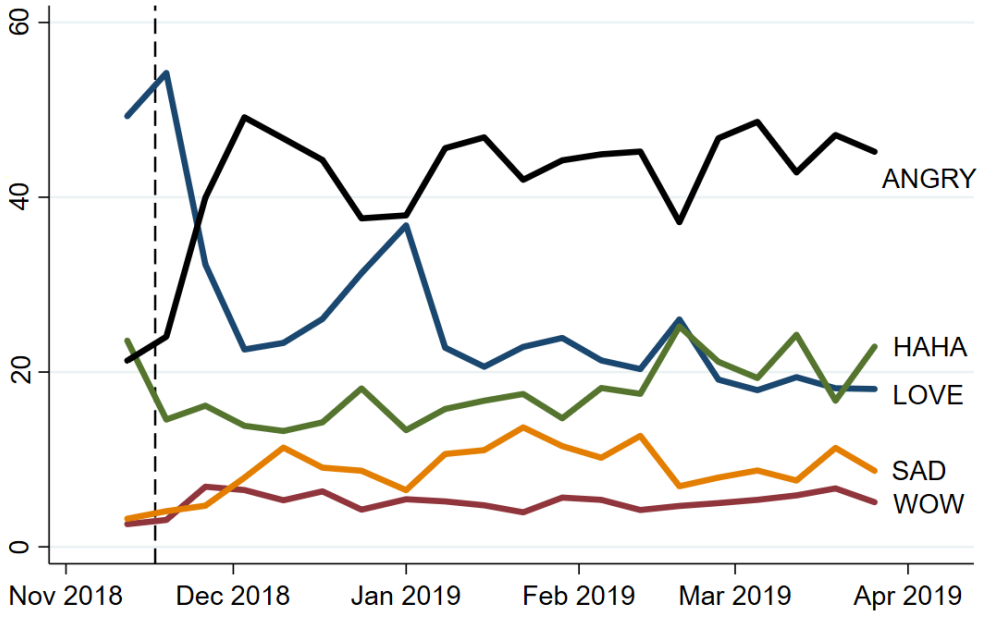
Notes: Sentences can be long and with many repetitions. For readability, we remove sequences of repeated tokens. The original phrases in French are in italics. Their English translation follows.

Figure E.3: Margins for Negative Sentiment Using TextBlob



Notes: This figure decomposes the increase in average negative sentiment using the method outlined in Section 3.3. We compute sentiment scores based on the TextBlob dictionary. Results are qualitatively similar to the main text results. 95% confidence intervals computed with the nonparametric bootstrap and 1000 iterations.

Figure E.4: Evolution of reactions



Notes: Weekly share of reactions to Facebook posts (in %). The dashed line corresponds to 11/17.

E.4 Political Partisanship Model

Our principal classification method is multinomial logistic regression. We consider the five largest French political parties: from right to left on the political spectrum, *le Rassemblement National* (RN), *les Républicains* (LR), *la République en Marche* (LREM), *le Parti Socialiste* (PS) and *la France Insoumise* (LFI). We parametrize the probability that a text snippet $\mathbf{x}$ is from party k as

$$\mathrm{P}(\text{party} = k \,|\, \mathbf{x}) = \frac{\exp(\mathbf{w_k} \cdot \mathbf{x} + b_k)}{\sum_j \exp(\mathbf{w_j} \cdot \mathbf{x} + b_j)},$$

in which $\mathbf{w_k}$ are specific coefficients to be estimated for party k. Given the large size of the vocabulary, we further penalize the multinomial logistic regression with the L2-norm (Ridge) to force some coefficients close to zero (Friedman, Hastie, Tibshirani et al., 2001). As some unigrams are not informative of political partisanship, the penalization mitigates over-fitting of the training set by shrinking coefficients.

There were very few far-right politicians (le RN) represented at the French Parliament in 2021, and the dataset of tweets only had 10,000 sentences for this party. To ensure a balanced dataset and estimate the model, we thus randomly draw 10,000 sentences from each party. We then shuffle the resulting corpus and split it into 80% training data and

20% test data. We build the classifier in the training set and evaluate its performance in the test set.

The model has accuracy, precision, and recall scores of 54-55%. A random guess would correctly infer the author's party 20% of the time. Our model thus assigns the correct party to a text snippet almost three times more often than a guess at random would. For comparison, Peterson and Spirling (2018) predict party affiliation with an accuracy between 60 and 80% for two parties. In this case, a guess at random would get the label right 50% of the time.

Table E.5 shows the model's confusion matrix, which suggests far-right and far-left speakers are slightly easier to predict than speakers from moderate parties. Table E.6 lists the most predictive words for each party according to our classifier. These words largely reflect each party's political stance. For instance, the Rassemblement National (RN) emphasizes words such as "immigration" and "islamism", whereas La France Insoumise (LFI) often mentions "protests" and "austerity". Figure E.5(a) presents the predicted partisanship of messages in our Facebook corpus for the first and the second scrape. Differences in the predicted partisanship of messages between both corpora are minimal.

Table E.5: Confusion Matrix of the Political Partisanship Model

		Predicted Party				
		RN	LFI	LR	LREM	PS
True Party	RN	0.63	0.11	0.09	0.11	0.07
	LFI	0.09	0.57	0.09	0.14	0.11
	LR	0.12	0.12	0.47	0.19	0.10
	LREM	0.08	0.11	0.15	0.53	0.13
	PS	0.07	0.11	0.11	0.17	0.53

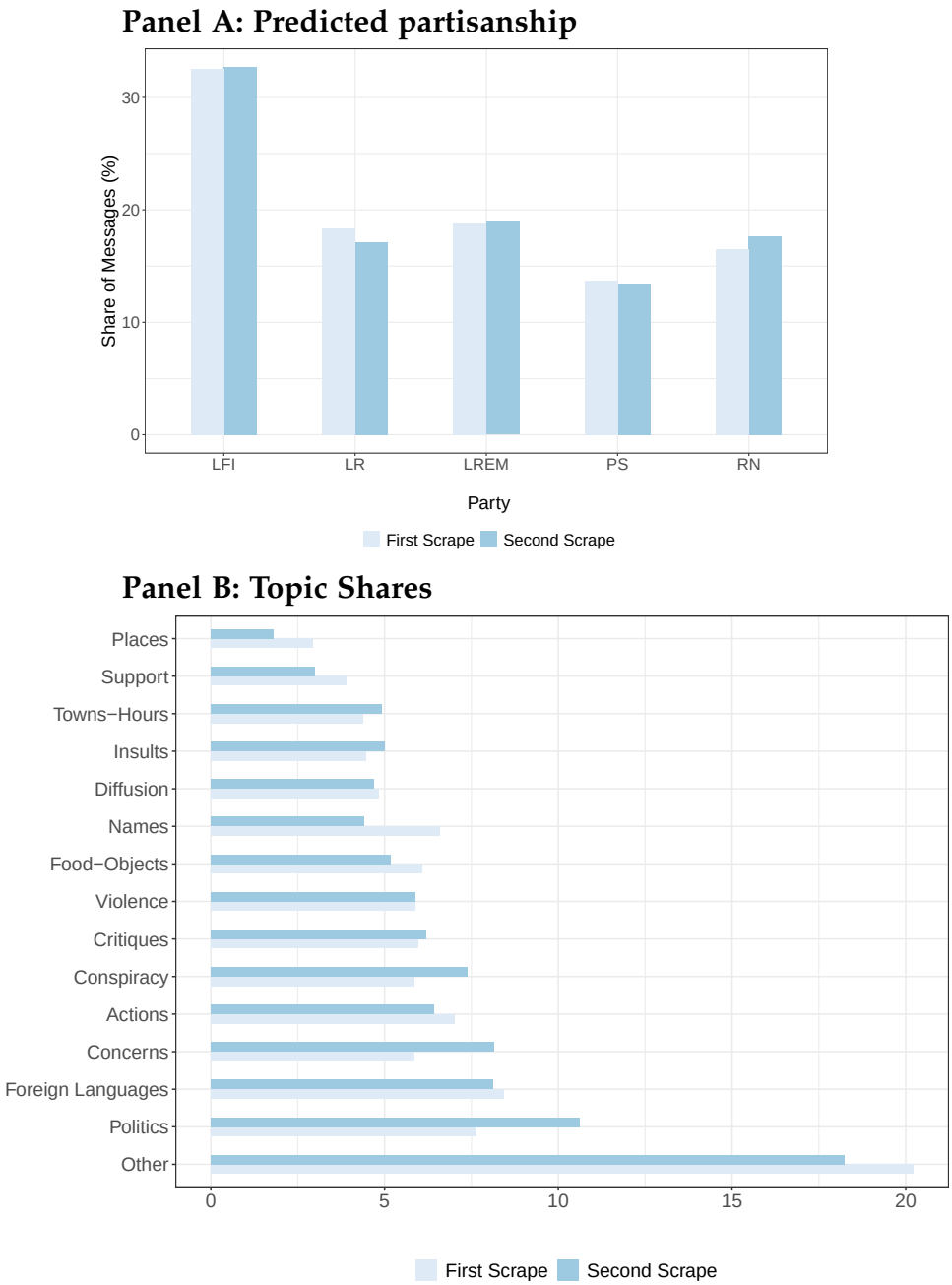
Notes: The confusion matrix C is such that C_{ij} is equal to the share of observations known to be of party i and predicted to be of party j.

Table E.6: Most Predictive Words Per Party

LFI	PS	LREM	LR	RN
insoumis	rabault	marcheur	peltier	mlp
insoumission	mans	denormandie	forissier	bardella
afcult	mayenne	adoption	vallee	gardois
larive	socialiste	larem	kuster	aliot
insoumise	94	complotisme	annemasse	marine
autain	mayennai	obstruction	restaurer	buissiere
incarcerer	riom	rencontr	lorion	bethune
planification	laval	avancee	ardechois	bruay
populaire	lacq	laureat	barnier	islamiste
toute	alfortville	gouffiercha	wauquiez	lievin
syndical	morancais	amont	loiret	compatriote
youtube	foncier	definition	ain	rachline
participez	apl	normandie	cesson	laxisme
autoritaire	jaures	charte	cope	rn
twitch	planete	integration	deficit	racaille
obono	cordialement	albi	dc	vardon
planifier	vallaud	avon	pris	riviere
foret	allocution	mobilite	nouzonville	soumission
manif	manceau	cluzel	savignat	perpignan
romainville	20e	stephanie	manipulation	ensauvagement
partagez	remuneration	bachelot	nicois	expulser
psychique	alimentation	grenoble	montargis	divergence
evasion	lamia	bachelier	exploiter	off
eau	schiappa	om	reconquerir	front
inutile	civique	intense	indefectible	ecrite
bolivie	ravie	contraception	ardeche	islamisme
programme	landes	incline	42	immigration
patricia	alim	gouvernance	briser	verlaine
degre	pdt	evoluer	fillon	frontiere
ivry	mayer	recette	ump	calai
rs	conciliation	attestation	fortement	immi
ecoeurer	fraternite	cohesion	evoque	beuvry
ariege	ivg	troll	echec	patriote
patissent	menetrol	croyance	democratiser	communique
mirepoix	clermont	durablement	lr	ravier
fac	lavalloi	chauny	larcher	clandestin
oms	herouville	habitation	bazin	insecurite
droite	unanimiter	menage	helas	incompetence
francis	applaudissement	apprenti	fur	bruaysiens
bifurcation	gauche	gouv	sociale	sketch
purificateur	bcp	inscription	lcp	philippot
repression	ba	approche	rythme	pas
muriel	acceleron	franc	ordinaire	ue
duplex	encommun	justifier	quentin	minier
austerite	inegalite	2025	poids	racaill
colonial	signent	rapp	oise	gafam
prive	mourenx	hydrogene	melange	juge
ressiguier	jospin	sejourne	progressisme	trahir
applaudir	insuffisanter	lune	race	banlieue
alternative	dividende	unanimite	archamp	auchel

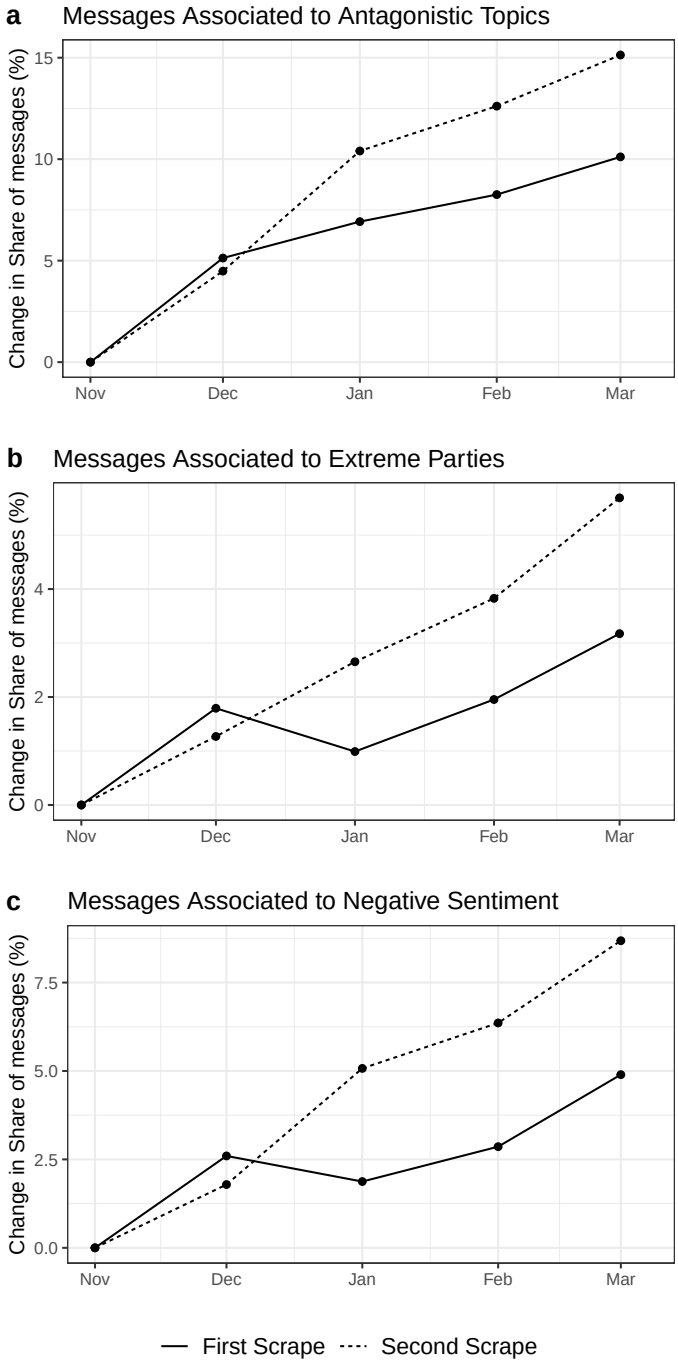
Notes: This table lists each party's 30 most predictive words according to our classifier. Words with large positive coefficients are most predictive of the speaker's party, so we simply rank the coefficients of words in descending order for each party to identify the top features.

Figure E.5: Partisanship and Topics for Each Data Collection



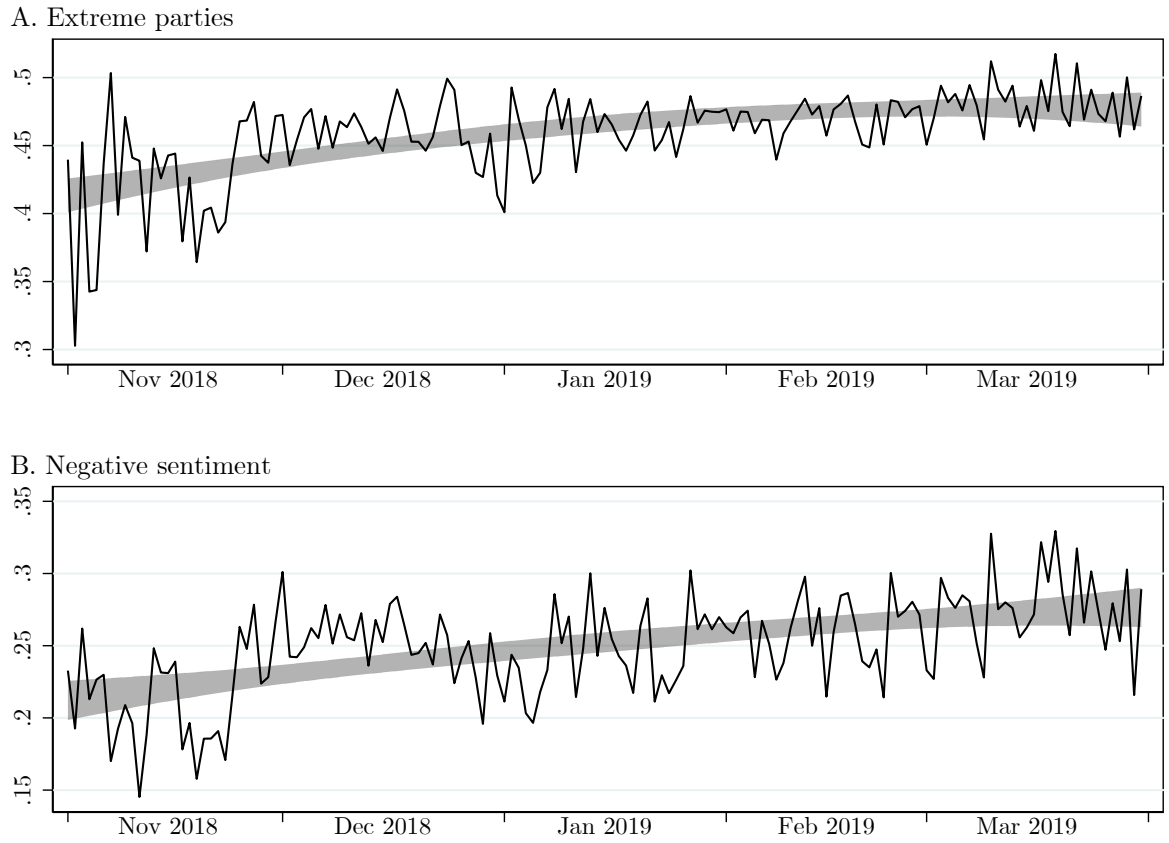
Notes: Panel a compares the predicted political leaning of sentences for the first (in light blue) and second (in dark blue) data collection. We assign a political leaning to each sentence in our corpus based on the probability of it being pronounced by a given party according to our supervised learning model. Panel b compares the share of messages assigned to each topic for our first (in light blue) and second (in dark blue) data collection on Facebook pages.

Figure E.6: Comparison of the Trends in Radical Attitudes for Each Data Collection



Notes: This figure compares observed trends in radical attitudes for our first (solid line) and second (dashed line) data collection on Facebook pages. Panel a presents changes in the share of sentences associated with an antagonistic topic. Panel b presents changes in the share of sentences associated with a politically extreme party (i.e., on the far left or the far right). Panel c presents changes in the share of sentences associated with negative sentiment.

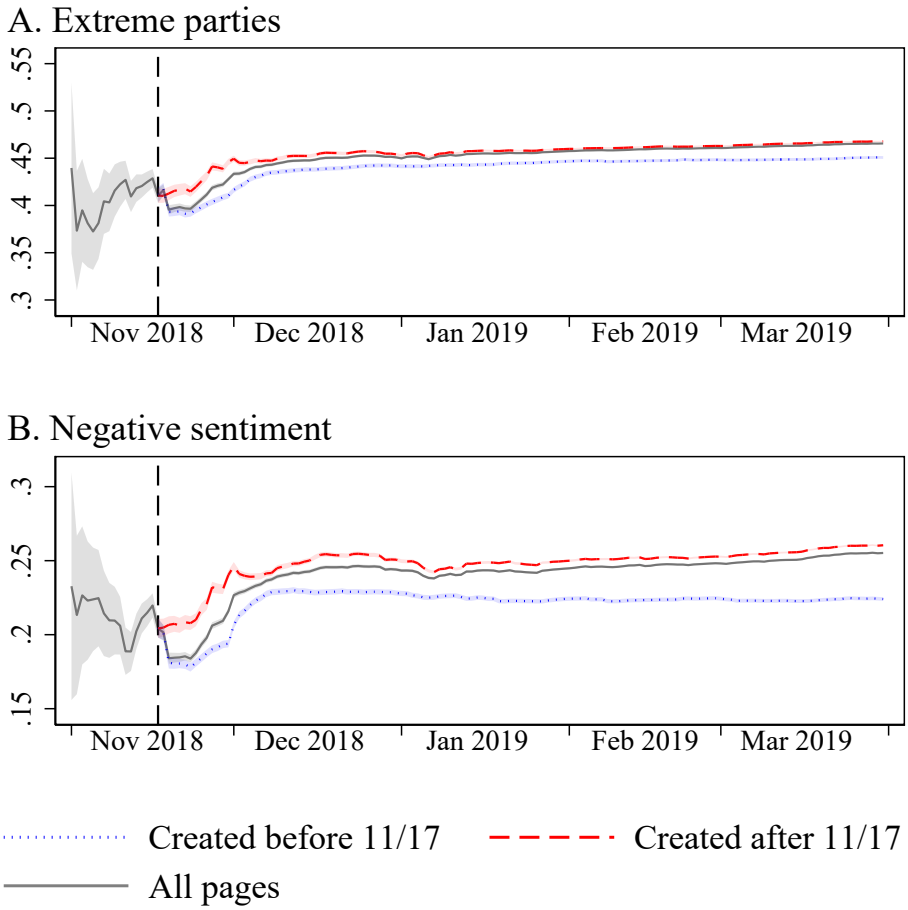
Figure E.7: Evolution of Online Violence: Robustness



Notes: This figure replicates Panel C of Figure 6 for two alternative measures of radicalism: the probability that the sentence was written by an affiliate of an extreme party (Panel A) and the probability that the sentence features negative sentiment (Panel B).

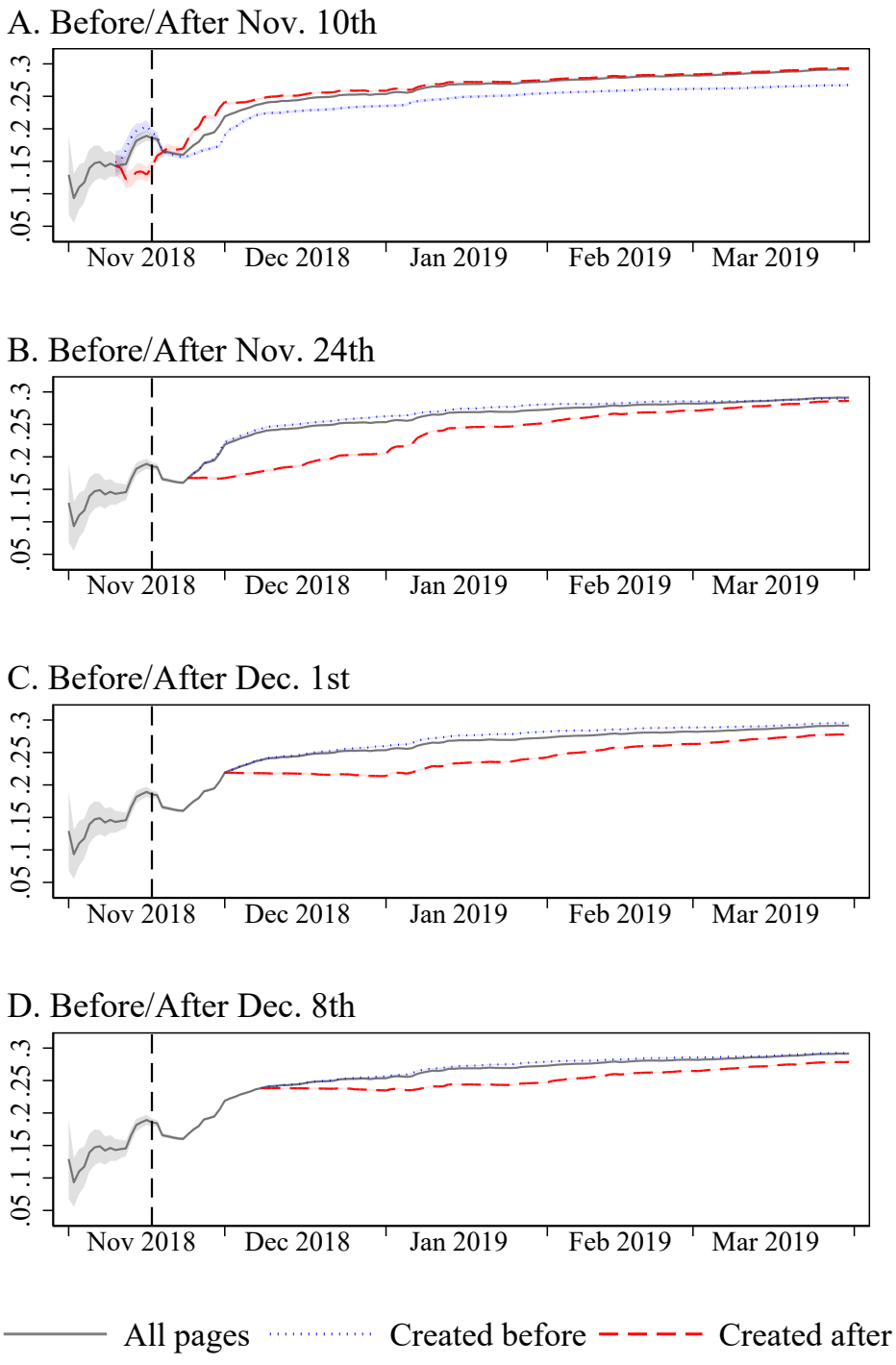
E.5 The revelation of a higher share of radicals after 11/17: robustness

Figure E.8: Bayesian updating on the share of radical discussants: alternative measures



Notes: This figure replicates the analysis described in Figure 7 for the share of messages associated with a politically-extreme party (Panel A) and the share of messages associated with negative sentiment (Panel B).

Figure E.9: Bayesian updating on the share of radical discussants: placebo dates

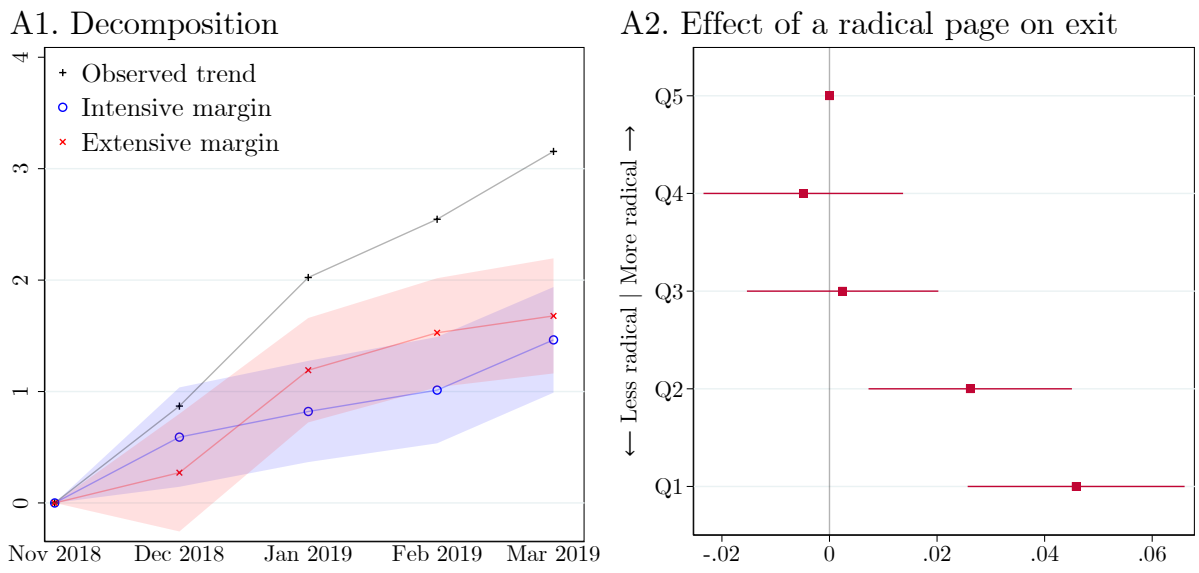


Notes: This figure replicates the analysis described in Figure 7 for four different cutoff dates.

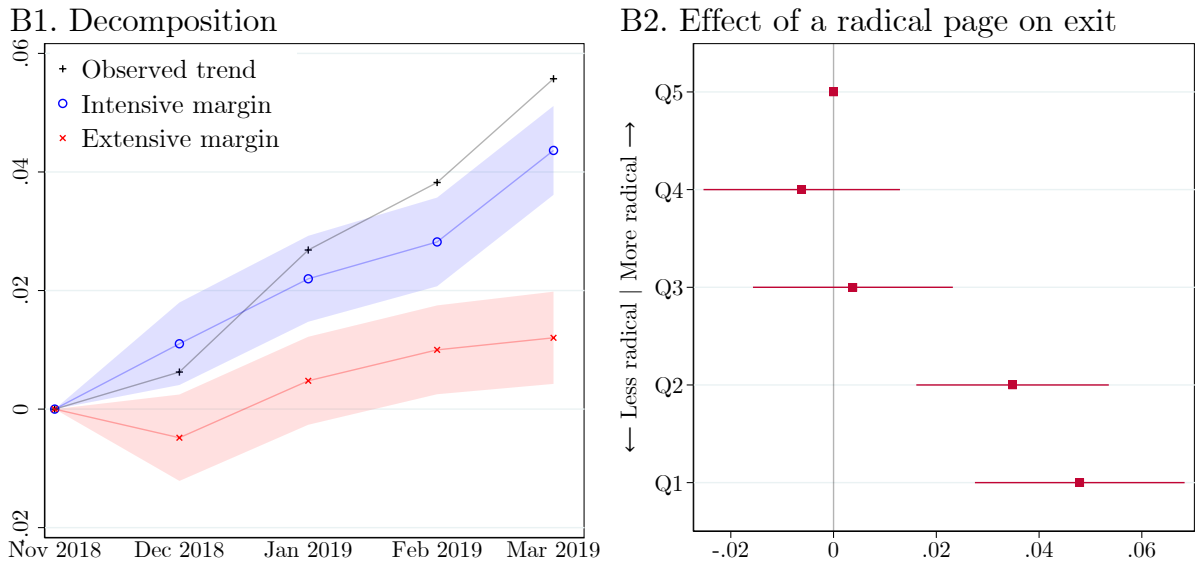
E.6 The crowd-out of moderate discussants: robustness

Figure E.10: The crowding-out of moderate online protesters: Alternative measures

A. Extreme parties



B. Negative sentiment



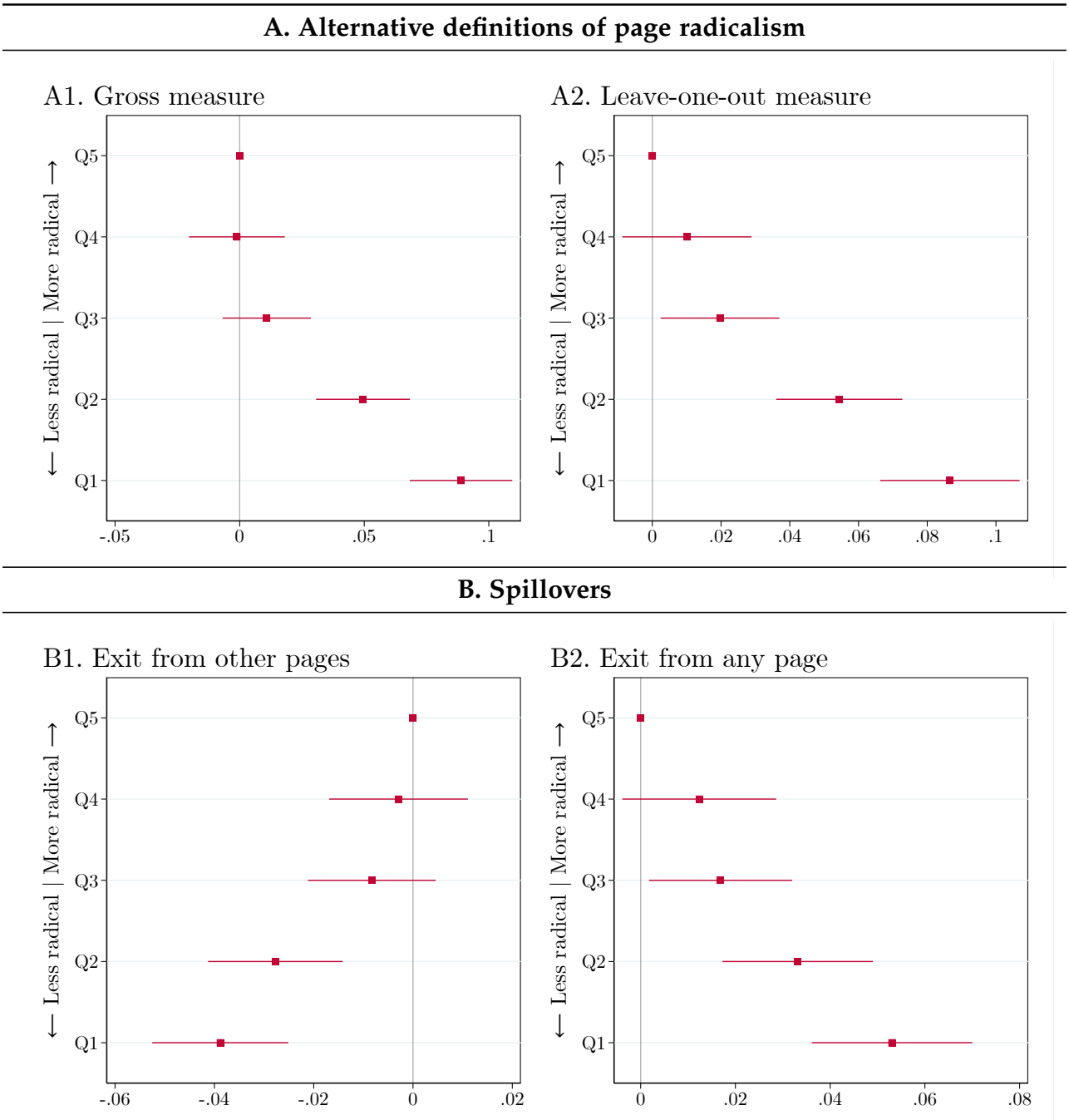
Notes: This figure replicates the results shown in Figure 8 for two alternative measures of radicalism: the probability that the sentence was written by an affiliate of an extreme party (Panel A) and (-1) times the sentiment score based on the Vader library (Panel B).

Table E.7: Page radicalization and the departure of the moderates: alternative specifications

	Dependent variable:				
	Probability of leaving the page				
	(1)	(2)	(3)	(4)	(5)
Panel A: Full Sample					
Moderate protester × Radical page	0.023***	0.023***	0.025***	0.049***	0.039***
	(0.003)	(0.003)	(0.003)	(0.006)	(0.007)
Observations	101,941	101,923	101,800	67,957	30,629
Number of discussants	57,897	57,881	57,852	24,076	10,025
Number of pages	359	341	325	292	265
R-Squared	0.02	0.11	0.13	0.58	0.60
Mean dependent variable	0.65	0.65	0.65	0.55	0.63
Panel B: Restricted sample					
Moderate protester × Radical page	0.010	0.017***	0.018***	0.045***	0.039***
	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)
Observations	30,629	30,629	30,629	30,629	30,629
R-Squared	0.04	0.14	0.17	0.53	0.60
Mean dependent variable	0.63	0.63	0.63	0.63	0.63
Month FE	✓	✓			
Page FE		✓			
Page-by-Month FE			✓	✓	✓
Discussant FE				✓	
Discussant-by-Month FE					✓

Notes: This table shows the OLS estimates of a regression of the probability of stopping posting on a Facebook page as a function of the interaction between the moderate dummy (having a fixed effect below the median of the distribution among discussants) and the (standardized) average discussant composition of the page measured at the sentence level for a given month. Radicalism is defined as the probability of posting a sentence associated with an antagonistic topic. Panel A shows estimation results on the full sample. Panel B shows estimation results on the most restricted sample corresponding to Column (5). We control for the main effects in the relevant specifications. In all specifications, we control for the number of sentences posted by the discussant on the page, by the number of sentences posted by the discussant on other pages, and by a binary variable indicating whether the discussant had posted on the page before. The sample is defined at the discussant-page-month level. We cluster standard errors at the discussant level. *: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$.

Figure E.11: Crowding-out over the distribution of protesters' radicalism: Robustness and extensions



Notes: This figure replicates Panel B in Figure 8 using different definitions of page radicalism (in Panel A) and different outcome variables (in Panel B). In Panel A1, we use the gross average of page radicalism at the sentence level $\mathbb{E}_{p,t}[Y]$ instead of the average of discussants' radicalism fixed effect associated with each sentence $\mathbb{E}_{p,t}[\delta]$ in Equation (7). In Panel A2, we use the leave-one-out average of discussants' radicalism fixed effect associated with each sentence $\mathbb{E}_{p,t,j\neq i}[\delta]$ in Equation (7) instead of the average. In Panel B1, the outcome variable is the probability to leave any other page the next month. In Panel B2, the outcome variable the probability to leave any page the next month.

E.7 The role of Facebook's algorithm

To study the impact of Facebook's algorithm on the radicalization of online mobilization, we take advantage of the structure of online discussions, which involve an initial post and its associated comments. While Facebook displays posts chronologically on Facebook pages, it does not deal with their associated comments similarly. Instead, undisclosed algorithms rank comments by what the platform calls "relevance." Since our dataset contains information on the ordering of comments shown to users at the time of the scrape, we can assess whether our radicalization measures are correlated with the recommendations of Facebook's algorithm.[8] To that end, we regress the rank of each comment in our text corpus on our measures of radicalism, controlling for a measure of the rank of the comment based on the time when the comment was posted. Rank measures are strongly positively correlated with each other, but the correlation is significantly lower than 1, which already suggests that Facebook alters the original ordering of comments..

Results are displayed in Table E.8. They show that comments associated with our radicalization measures are more likely to be found higher on the list. For example, Column (1) in Panel A shows that comments associated with antagonistic topics are displayed at a rank 14% higher than other comments. The same patterns appear if we focus on the probability of being a "star comment", which we take as one of the first four comments below the post. Such comments are likely to appear in users' newsfeeds without further clicking and are, therefore, much more likely to be salient and read by users. Column (1) in Panel B shows that messages featuring a negative sentiment are 0.9 p.p. more likely to belong to this selected set, which corresponds to a 9% increase in the baseline probability. These results show that a chronological order of comments would have provided discussants with less radical content.

We assess the robustness of these results to several concerns. First, since posts vary a lot in their content and the number of comments they generate, we also control for post fixed effects in the other columns of the table. Column (2) shows that the results are still sizable if we use post fixed effects. For example, if a sentence belongs to the three radicalism categories (8% of the full sample), our estimates in column (2) of Panel A show that its rank is, on average, 16% higher than a sentence that does not belong to either category (32% of the full sample). Second, some posts are made of several sentences, which may bias the results if Facebook's algorithm treats posts of different length

[8]The Facebook account that we created to scrape this data was historyless, hence unlikely to affect Facebook's recommendation algorithm.

Table E.8: Comments' Rank and Radical Content

Panel A	**Dependent variable:**			
	Rank of the comment (in log)			
	(1)	(2)	(3)	(4)
Antagonistic Topic	-0.136***	-0.081***	-0.079***	-0.033***
	(0.006)	(0.004)	(0.004)	(0.006)
Extreme Parties	-0.046***	-0.017***	-0.020***	-0.029***
	(0.004)	(0.003)	(0.003)	(0.005)
Negative Sentiment	-0.136***	-0.065***	-0.112***	-0.043***
	(0.007)	(0.004)	(0.005)	(0.006)
Mean dependent variable	4.462	4.480	4.965	3.090
R-Squared	0.713	0.813	0.843	0.812
Panel B	**Dependent variable:**			
	Comment is among the first four (in %)			
	(1)	(2)	(3)	(4)
Antagonistic Topic	0.339***	0.309***	0.145***	0.679***
	(0.076)	(0.042)	(0.046)	(0.193)
Extreme Parties	0.437***	0.190***	0.167***	0.274
	(0.052)	(0.034)	(0.038)	(0.176)
Negative Sentiment	0.897***	0.340***	0.262***	0.723***
	(0.081)	(0.046)	(0.051)	(0.198)
Mean dependent variable	10.547	10.171	7.130	16.716
R-Squared	0.248	0.480	0.468	0.570
Post Fixed Effect		✓	✓	✓
Single-sentence Posts			✓	✓
User Fixed Effect				✓
Observations	1,889,894	1,881,976	1,133,399	177,283

Notes: This table shows estimates of OLS regressions at the sentence level. We restrict the text corpus to comments (and exclude original posts). In Panel A, the dependent variable is the (log) rank of the comment suggested by Facebook at the time of the scrape. In Panel B, the dependent variable is a dummy variable equal to 1 if the comment is among the first four comments suggested by Facebook at the time of the scrape. "Antagonistic Topic" is a dummy variable equal to 1 if the sentence is classified as belonging to an antagonistic topic. "Extreme Parties" is a dummy variable equal to 1 if the sentence is attributed to an extreme party. "Negative Sentiment" is a dummy variable equal to 1 if the sentence is associated with a negative sentiment value. In all specifications, we control for the counterpart of the dependent variable, based on chronological order. In Columns (3)-(4), we restrict the sample to single-sentence comments. In Column (4), we control for user fixed effects using information from our second scrape. In all regressions, we cluster standard errors at the post level. *: p<0.01, **: p<0.05, ***: p<0.1.

differently and the length of radical posts differs from that of other posts. However, Column (3) shows that the results are similar if we restrict the sample to single-sentence posts. Finally, one could think that the algorithm does not highlight radical sentences, but simply sentences made by popular discussants. This effect would bias our results if popular discussants were more likely to post radical content. However, Column (4) shows that our results are robust to controlling for discussant fixed effects.

References

Arora, Sanjeev, Yingyu Liang, and Tengyu Ma, "A simple but tough-to-beat baseline for sentence embeddings," *Conference paper at ICLR 2017*, 2017.

Demszky, Dorottya, Nikhil Garg, Rob Voigt, James Zou, Matthew Gentzkow, Jesse Shapiro, and Dan Jurafsky, "Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings," in "Proceedings of NAACL-HLT" 2019, pp. 2970–3005.

Friedman, Jerome, Trevor Hastie, Robert Tibshirani et al., *The Elements of Statistical Learning*, Vol. 1, Springer series in statistics New York, 2001.

Leroy, Claire, "Raising Take-up of Welfare Programs: Evidence from a Large French Reform," 2024. Mimeo CREST.

Peterson, Andrew and Arthur Spirling, "Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems," *Political Analysis*, 2018, *26* (1), 120–128.

Rieder, Bernhard, "Studying Facebook via Data Extraction: The Netvizz Application," in "Proceedings of the 5th annual ACM web science conference" ACM 2013, pp. 346–355.

Yan, Xiaohui, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng, "A Biterm Topic Model for Short Texts," in "Proceedings of the 22nd international conference on World Wide Web" 2013, pp. 1445–1456.