The Lifecycle of Protests in the Digital Age

Pierre C. Boyer Germain Gauthier Yves Le Yaouanq

Vincent Rollet Benoît Schmutz-Bloch

December 2024

We propose a theory of protest dynamics with heterogeneous protest technology and intensity. The ability to mobilize online reduces the likelihood of coordination failures at both the extensive (engagement) and intensive (violence) margins. We build a dynamic coordination game with strategic substitutability and endogenous learning, and use it to characterize a *crowd-in-then-crowd-out* sequence in which social media initially helps launch massive protests, but then encourages radical factions to turn violent, leading moderates to leave the movement. This sequence is illustrated using online and offline data on the 2018 Yellow Vest uprising in France, whose early success and popularity were abruptly undermined by street violence. First, spatial regressions confirm that online and offline mobilizations reinforced each other at the beginning of the movement. Second, our textual analysis reveals that online conversations among protesters progressively radicalized. Using a decomposition with discussant, page, and period fixed effects, we show that (i) half of this trend was due to changes in the composition of online protesters and (ii) more radicalized pages drove out moderate discussants.

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, 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. 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.
"I have a Dream" by Martin Luther King (August 28th, 1963)

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 largely peaceful, others 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 interplay between the likelihood, size, intensity, and persistence of protests. We propose a model of protest dynamics based on coordination failures and information acquisition that helps to understand when protests can emerge and thrive, but also why they can radicalize and eventually die out. Our framework is set in the context of contemporary protest movements that combine traditional street protests with online mobilization through social media. While social media is known to facilitate the emergence of protests (Zhuravskaya, Petrova and Enikolopov, 2020; Aridor, Jiménez-Durán, Levy and Song, Forthcoming), its ability to quickly organize massive protests may also contribute to making these protests vulnerable in the long run (Tufekci, 2017). We show that the impact of social media on protests can change from positive to negative over time, a pattern we illustrate by combining online and offline protest data.

We follow the tradition of modeling protests as a game in which payoffs are determined by the total number of players who choose to mobilize. To model protest intensity, we also posit that mobilization can be either peaceful or violent. Protests are thus characterized by two common features: their size and their level of violence. This combination gives rise to four types of protests (labeled in the model as *routine*, *rally*, *riot*, and *revolution*) that can be intuitively mapped to real-world situations. Whether one type emerges as the equilibrium depends on the degree of strategic complementarity and substitutability between violent and peaceful protesters: while violent protesters value all protesters positively, peaceful protesters dislike their violent counterparts. For simplicity, we consider only three types of players: they can be either passive and never mobilize, moderate, or radical. Moderates are responsible for the *extensive* margin of mobilization (mobilize or not mobilize) and radicals are responsible for the *intensive* margin (mobilize peacefully or violently).

Players face uncertainty about their respective shares of the population and update their beliefs about these shares using information from past protests. Some equilibria

imply an identification problem, making *learning traps* possible even with an infinite number of periods (Fudenberg and Levine, 1993). These traps can affect the extensive margin (if players overestimate the share of passives), the intensive margin (if players misperceive the share of radicals), or both margins simultaneously. We introduce social media by letting players in each period also decide whether to participate in *online* protests before making their *offline* protest decision. Online protests are less costly, so social media reduces the occurrence of learning traps. In particular, social media can help launch protests that would never have started otherwise because players were overestimating the share of passives in the population. However, social media reduces the occurrence of all learning traps, including those that prevent radicals from coordinating on violent action.

Equipped with this framework, we then focus on the classic situation where social media has helped a protest movement start and gain initial momentum, and we characterize the conditions under which this movement will persist over time and in what form. In particular, we fully characterize the preferences and beliefs of the population that will lead to a *crowd-in-then-crowd-out* sequence in which this early reliance on social media will also cause the movement to radicalize and fade away, converging on a combination of violent street riots and violent online discussions, without popular support. This sequence sheds light on several paradoxes: For example, why low participation costs do not always make protests more sustainable; or why state repression in the form of Internet shutdowns during social unrest can sometimes backfire (Rydzak, Karanja and Opiyo, 2020). Interestingly, it does not require any technological bias of social media in favor of radical content: such a bias, if present at the outset, might even prevent social media from playing its catalytic role, leaving the movement stillborn.

In the second part of the paper, we establish the empirical relevance of the *crowd-in-then-crowd-out* sequence in the context of The Yellow Vest movement, one of the most significant episodes of social unrest in recent French history, which also shared many characteristics with concurrent protest movements around the world.[1] Sparked by an

[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). More generally, the movement may be associated with the return of local politics that has been documented all over the world (Della Porta and Diani, 2020; Le Galès, 2021).

online petition 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. Consistent with previous research in other settings, we first show that early online activity was highly predictive of the occurrence of a roadblock on 11/17. We then describe a lesser-known phenomenon in the literature: the initial street protests triggered a second wave of online activity in the weeks that followed. This second wave, which is directly observable in daily time series, was almost fully concentrated in municipalities that had mobilized on 11/17. Such a rebound effect is consistent with our way of modeling *crowd-in*, in which social media not only helps to launch protests, but also helps young protests gain initial momentum. It suggests that the 11/17 protests, which were quite successful and under intense media scrutiny, helped spread information about the popularity of the Yellow Vests, which increased the intensity of subsequent online mobilization, thereby triggering a positive feedback loop between early online and offline mobilization. Both directions of this loop are further confirmed using two different instrumental variable strategies based on the progressive deployment of the 4G network and local variation in the density of roundabouts.

Despite this online-offline feedback loop, however, the 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. Using our municipal dataset, we first show descriptively that more violent street protests in 2018 were associated with the subsequent formation of smaller protest communities, both online and offline. Then, to assess whether this shrinking pattern was indeed driven by the departure of moderate protesters, we leverage another dataset of discussions on Yellow Vest Facebook pages, for which we can track individual protesters' comments over time. Using various 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 antagonistic, negative, and politically polarized.

We exploit our panel dataset to decompose the radicalization process into an ex-

tensive margin (changes in the composition 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 *crowd-out* of moderate discussants by more radical ones at the aggregate level. We further show that moderate discussants left Facebook pages where discussions had become more radical. This effect is quantitatively important and is robust to controlling for discussant fixed effects that account for the sorting of discussants across pages, and even discussant-by-period fixed effects, which account for the entire mobilization history of discussants.

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

[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.

learn about it by observing data from past protests. The idea that protesting decisions 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 that, for some plausible values of the population's preferences, the strategic interaction between moderates and radicals implies that the dynamics of protests display 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 Vests movement. A similar dynamic arises in the models by Correa (2022) and Enikolopov, Makarin, Petrova and Polishchuk (2020b), but for different reasons. In Correa (2022), 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 and show that incentives to attend a protest once have dynamic consequences if a significant share of one's social network also turns out.[6]

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

[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).

[6] Ives and Lewis (2020) study empirically the conditions under which peaceful protests ("rallies") turn into "riots", and Alsulami, Glukhov, Shishlenin and Petrovskii (2022) analyze the dynamics of a mathematical (non-economic) model of differential equations, which they apply to the Yellow Vests movement.

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 emergence of protest movements, with most studies finding a positive effect (e.g., Acemoglu, Hassan and Tahoun, 2018; Larson, Nagler, Ronen and Tucker, 2019; Enikolopov, Makarin and Petrova, 2020a; Fergusson and Molina, 2021).[7] 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).[8] 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 descriptive evidence (Bastos, Mercea and Charpentier, 2015).

Finally, we discuss why social media can also contribute to the premature demise of protest movements, consistently with Tufekci's (2017) insights. To that end, we link the issue of violence to the use of social media, which have long been accused of fostering ideological segregation through filter bubbles and echo chambers (Pariser, 2011).[9] We contribute to this debate by introducing the paradoxical result that an algorithmic bias towards violent discussions may not necessarily lead to more violent protests be-

[7]Other studies focus on different outcomes, such as hate crimes (Bursztyn, Egorov, Enikolopov and Petrova, 2024) or voting behavior (Madestam, Shoag, Veuger and Yanagizawa-Drott, 2013; Fujiwara, Muller and Schwarz, 2024).

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

[9]Several studies have provided experimental evidence that social media use is indeed associated with political polarization, but through conflicting mechanisms. For example, Levy (2021) shows that Facebook's algorithm is less likely to expose users to posts from news outlets with opposing views, while doing so would reduce their negative attitudes toward 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. Overall, how and to what extent social media affects political polarization is still debated (see, e.g., Ross Arguedas, Robertson, Fletcher and Nielsen, 2022).

cause of its contradictory effects on the different factions behind the movement. We also contribute to the methodological toolkit of this literature by proposing several methods to measure radicalization and its mechanisms, taking advantage of the structure and content of social media data.

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

In this section, we present a dynamic model of political participation based on strategic uncertainty and information revelation, where social media has non-trivial effects on protests dynamics.

2.1 The protest game

Our framework involves repeated protest participation decisions. We start by describing and analyzing the stage game of offline protests in the absence of social media.

Preferences. We consider a population of agents of mass one. Each agent is characterized by a fixed type $\theta \in \mathbb{R}$ that measures the willingness to participate in the protest movement.[10]

Participation decisions take three possible values, $a = 0$ (not participating), $a = 1$ (participating in a peaceful manner), and $a = v > 0$ (participating in a violent manner).[11] The utility from not participating is normalized to zero. Preferences depend on five parameters α, β, γ, $\underline{c}$ and $\overline{c}$ where α measures the value of a peaceful protest, β and γ measure the gain or loss from violence, and $\underline{c}$ and $\overline{c} > \underline{c}$ measure the direct cost of

[10] This model is consistent with an interpretation of θ as reflecting a protester's expressive concern, or her 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.

[11] Although violence will play a signaling role in the dynamic model, protesters do not resort to violence 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.

peaceful and violent protest, respectively. Participation decisions depend on the mass of individuals selecting either type of action: An individual i of type θ_i who plays $a_i = 1$ reaps a payoff equal to

$$U[a_i = 1, \{a_j\}] = \theta_i + \alpha \mathbb{E}\mathbb{1}_{a_j=1} - \beta \mathbb{E}\mathbb{1}_{a_j=v} - \underline{c} \tag{1}$$

while the same individual playing $a_i = v$ receives a payoff

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

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 differences in experienced utilities from participating depending on the size of the crowd. In addition, more extreme types have a greater gain from choosing violence.

We assume $\alpha, \beta, \gamma > 0$, which implies that all interdependencies take the form of a strategic complementarity, except that violent action discourages peaceful protests.[12] We also assume $\gamma > \alpha$: complementarities are stronger for violent than for peaceful protests.

Types and uncertainty. Agents' preferences θ are heterogeneous. A fraction $1 - \mu$ is *passive* ($\theta = \theta_P \approx -\infty$) and plays $a_P = 0$. 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 restrict the analysis to situations where radicals never abstain ($a_R \in \{1, v\}$) and moderates never engage in violent action ($a_M \in \{0, 1\}$).[13]

The parameters λ and μ are uncertain. In the dynamic version of the game, from subsection 2.2 onwards, the population uses information about past protests to update its 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 that all groups share a common belief. In the stage game, we capture the populations' beliefs via the expectations $\mathbb{E}[\lambda\mu]$ and $\mathbb{E}[(1-\lambda)\mu]$ of the share of radical and moderate individuals, respectively.

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

[13] For example, this will be the case if $(v+1)\theta_M + \gamma < \overline{c}$ and $\theta_R > \underline{c}$. This restriction puts the emphasis on situations of social unrest, where the population is prone to mobilizing. Allowing radicals to play $a_R = 0$ would not affect our main results. In Section 2.2, we allow for this extension to study short-term dynamics.

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 and μ. The stage game typically admits multiple equilibria. To limit the number of cases to consider, we focus on *Strong Nash Equilibria* (Aumann, 1959), that is, equilibria where no individual or coalition can profitably deviate.[14] One possible justification for using this equilibrium concept is that political factions have the ability to coordinate on their preferred action among those that define a Pareto-undominated Nash equilibrium.[15]

Equilibria. An equilibrium is described by a pair (a_M, a_R), where $a_M \in \{0, 1\}$ is the strategy of the moderates and $a_R \in \{1, v\}$ is that of the radicals. We interpret the four possible equilibria as follows: the $(0, 1)$ equilibrium describes a *Routine* situation in which both types choose their default action; conversely, if moderates join radicals, they form a *Rally*—equilibrium $(1, 1)$. Radicals, however, may choose to protest violently. If they do so without the support of moderates, they lead a *Riot*—equilibrium $(0, v)$, but if moderates protest peacefully alongside them, the situation amounts to a *Revolution*—equilibrium $(1, v)$. 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. We use Lemma 1 to solve the model.

Lemma 1 *The equilibria of the stage game are described on Figure 1 in the (θ_M, θ_R) plane, and fully characterized by the thresholds $\overline{\theta}_R$, θ^*_R, $\underline{\theta}_R$, $\overline{\theta}_M$, and $\underline{\theta}_M$ defined as follows:*

$$\begin{cases} v\overline{\theta}_R = \overline{c} - \underline{c}, \\ v\theta^*_R + \gamma\mathbb{E}[\lambda\mu] - \alpha\mathbb{E}[\mu] = \overline{c} - \underline{c} \text{ if } \gamma\mathbb{E}[\lambda\mu] > \alpha\mathbb{E}[\mu], \\ v\underline{\theta}_R + (\gamma - \alpha)\mathbb{E}[\lambda\mu] = \overline{c} - \underline{c}, \\ \overline{\theta}_M + \alpha\mathbb{E}[(1-\lambda)\mu] - \beta\mathbb{E}[\lambda\mu] = \underline{c}, \\ \underline{\theta}_M + \alpha\mathbb{E}[\mu] = \underline{c}. \end{cases}$$

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 $\overline{\theta}_M > \underline{\theta}_M$, illustrating that violent movements crowd out peaceful participation. The position of θ_R relative to: (i) $\overline{\theta}_R$ determines

[14] Strong Nash Equilibria do not always exist, but they do in our setting. This criterion implies that every surviving equilibrium is Pareto-efficient.

[15] Whenever possible, we break remaining ties in favor of the equilibrium with the lowest participation on either margin. We focus on the interior of each region in the main text and treat the frontiers in Appendix A.1.

whether playing $a_R = v$ is a dominant strategy for radicals; (ii) $\underline{\theta}_R$ determines whether radicals prefer the equilibrium $(0,1)$ over $(0,v)$ or vice versa. Finally, there may be situations where radicals prefer the equilibrium $(0,v)$ over $(1,1)$, even though playing $a_R = v$ is not a dominant strategy for them. This situation will happen if $\theta_R > \theta_R^*$, but only in the case where the complementarity of violent protests outweighs that of peaceful protests ($\gamma\mathbb{E}[\lambda\mu] > \alpha\mathbb{E}[\mu]$), so that $\theta_R^* < \overline{\theta}_R$. When $(\theta_M, \theta_R) \in (\underline{\theta}_M, \overline{\theta}_M) \times (\theta_R^*, \overline{\theta}_R)$, our equilibrium selection leaves some indeterminacy: both $(0,v)$ and $(1,1)$ are Pareto-undominated Nash equilibria, the former is preferred by the radicals and the latter by the moderates.

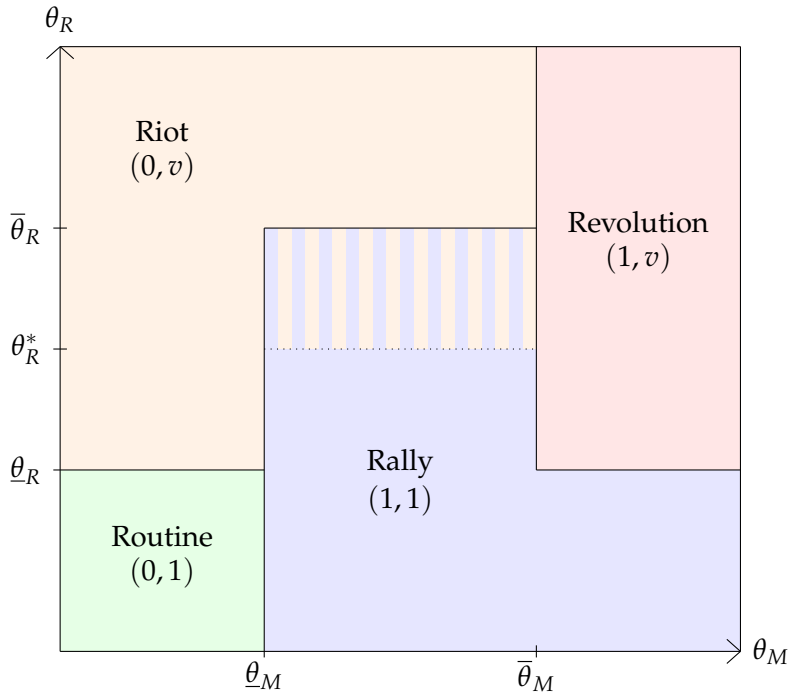
The strategic impact of moderates' preferences on radicals' protests is non-monotone. To see this, consider the region where $\theta_R \in (\underline{\theta}_R, \theta_R^*)$. When $\theta_M < \underline{\theta}_M$, moderates are never active, and radicals resort to violent action. The same is true when moderates are always active ($\theta_M > \overline{\theta}_M$). For intermediate values of θ_M, radicals remain peaceful, so as not to exclude moderates from the movement.[16]

Comparative statics. As a preliminary to the analysis of the role of people's beliefs about the population's preferences 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, and θ_R^* upwards, as illustrated in Panel A of Appendix Figure A.1. Moderate players become more prone to participation, regardless of the action chosen by radical players. If the type of radical players is very low or very high ($\theta_R \notin [\underline{\theta}_R, \overline{\theta}_R]$), this increase does not affect their behavior. Conversely, in the intermediate case, an increase in the share of moderates has an ambiguous effect on the behavior of radicals depending on the baseline level of participation: it can "pacify" a fringe riot (from $(0,v)$ to $(1,1)$ in the region $[\underline{\theta}'_M, \underline{\theta}_M]$) through the increased participation of moderates; conversely, it can radicalize a large peaceful movement (from $(1,1)$ to $(1,v)$ in the region $[\overline{\theta}'_M, \overline{\theta}_M]$), because radicals can now play $a_R = v$ without fearing that moderates will leave the movement.

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 $\underline{\theta}_R$ and θ_R^* downwards and $\overline{\theta}_M$ upwards, as shown in Panel B of

[16] This pattern echoes the results of Bueno de Mesquita (2013), who shows that violence can only occur in situations of intermediate hardship because widespread poverty convinces everyone to join a large and peaceful movement, whereas prosperity discourages everyone from mobilizing. In our model, a joint increase in θ_M and θ_R can increase the likelihood of violence, from $(0,1)$ to $(0,v)$ and from $(1,1)$ to $(0,v)$ or $(1,v)$, or decrease it, from $(0,v)$ to $(1,1)$.

Figure 1: Equilibria of the Stage Game: the 4R of Revolts



Notes: The striped area corresponds to the equilibrium $(1, 1)$ if $\gamma \mathbb{E}[\lambda\mu] < \alpha \mathbb{E}[\mu]$. In that case, θ_R^* is not defined. Conversely, if $\gamma \mathbb{E}[\lambda\mu] > \alpha \mathbb{E}[\mu]$, the equilibrium is indeterminate, with moderates preferring $(1, 1)$ and radicals preferring $(0, v)$.

Appendix Figure A.1. This variation does not change the action chosen by moderates and has an ambiguous effect on radicals' behavior. Indeed, it might lead them to start protesting violently (e.g., switch from $(0, 1)$ to $(0, v)$ or from $(1, 1)$ to $(1, v)$), but it might paradoxically pacify a violent movement (e.g., from $(1, v)$ to $(1, 1)$). In the latter case, this is because radical players, now more numerous, must refrain from violent action for fear of moderates leaving the movement.

2.2 Dynamics of protests without social media

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 equilibrium framework. For simplicity we abstract from modeling the response of the government, which could affect protest dynamics. In Section 2.4 we consider the strategic response of a government contemplating a shutdown of social media.

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.[17]

All players start the game with a common prior belief over (λ, μ) described by the full-support pdf $\chi_0 : [0,1]^2 \to [0,1]$.[18] 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 do not resolve the indeterminacy between $(0, v)$ and $(1, 1)$ in the region where multiple equilibria are allowed (the striped area in Figure 1), but instead assume that the same equilibrium is played every time these equilibria co-exist. This can reflect that one of the two political groups has a higher ability to coordinate and impose its preferred equilibrium on the other.

Given this selection rule, we write $a^*(\chi) = [a_M^*(\chi), a_R^*(\chi)]$ for the equilibrium of the stage game under belief χ.

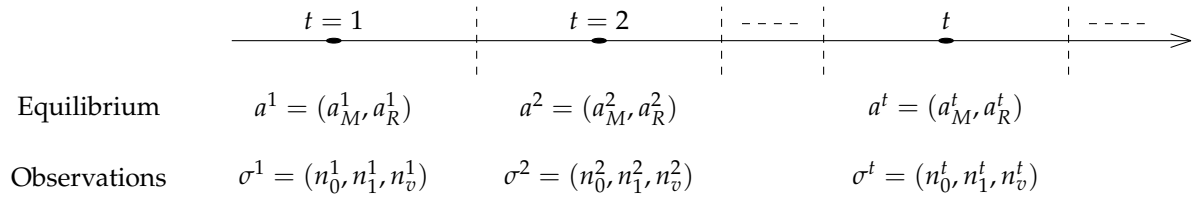
Information. After each date t, the behavior of n players at the last stage game is publicly displayed. These n players are randomly, uniformly and independently selected from the population. That is, the probabilities with which a selected individual is passive, moderate or radical equal $1 - \tilde{\mu}$, $(1 - \tilde{\lambda})\tilde{\mu}$ 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 $a^s = (a_M^s, a_R^s) \in \{0,1\} \times \{1, v\}$; (ii) the number n_a^s of individuals playing action $a \in \{0, 1, v\}$ at date s, where $n_0^s + n_1^s + n_v^s = n$. We write $\sigma = (n_0, n_1, n_v)$ 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.

[17] Alternatively, players can be patient provided the benefits and costs of political participation are time-separable (as under discounted expected utility maximization). In that case, the equilibria also define a Perfect Bayesian Equilibrium of the dynamic game.

[18] 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.

Figure 2: Timeline



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 were informed about $\tilde{\lambda}$ 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.

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.

Preliminary observations. To see why learning traps might arise and why they give rise to inefficiencies, consider the case where the full-information equilibrium is $(1,1)$, but the protesters' beliefs underestimate μ. This misperception prompts them to play $(0,1)$, which, over time, reveals the share of radicals $\tilde{\lambda}\tilde{\mu}$ perfectly but leaves uncertainty about the share of active players $\tilde{\mu}$. If initial pessimism was strong enough, it is possible that $(0,1)$ is played forever instead of the Pareto-dominant equilibrium $(1,1)$.

The profile $(1,v)$ cannot be played in a learning trap, as it would reveal the values of $\tilde{\lambda}$ and $\tilde{\mu}$ perfectly, yielding rational expectations. Conversely, if the full-information equilibrium is $(0,1)$, then it is reached with probability one from any correctly specified prior. Learning traps are, therefore, asymmetric: protesters might fail to start a movement that would have been successful under full information, but an unpopular social protest is never artificially maintained. Proposition 1 summarizes all possibilities.

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

(i) those that reduce the extensive margin of protests;

(ii) those that reduce the intensive margin of protests;

(iii) those that affect both margins in opposite ways by transforming a riot $(0,v)$ *into a rally* $(1,1)$ *or vice versa.*

Table 1 describes all possible learning traps by: (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. The first three rows of Table 1 confirm that information frictions can systematically hinder the coordination that is necessary to give rise to large-scale protests. In the first two cases, this happens because moderates systematically underestimate their share, and thus their payoff to protesting. In the third case, the intensive margin is lower than under full information because radicals misperceive their share. This learning trap can occur when the population underestimates $\tilde{\lambda}$, which is intuitive, but also when it overestimates it: in that case, radicals refrain from violent action for (unfounded) fear of excluding moderates from the movement. The last two rows of Table 1 reveal that

Table 1: List of learning traps

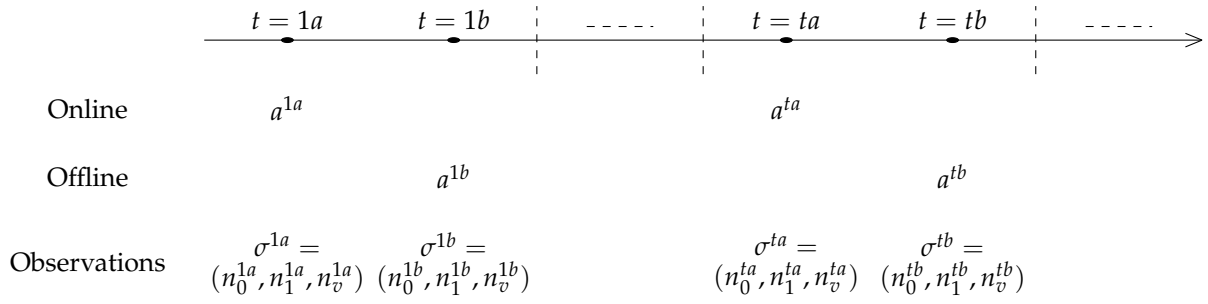
Margin affected	Self-confirming equilibrium	Full-information equilibrium	Long-run beliefs
Extensive	$(0,1)$	$(1,1)$	$\mathbb{E}_\chi[\lambda\mu] = \tilde{\lambda}\tilde{\mu}$, $\mathbb{E}_\chi[\mu] < \tilde{\mu}$
	$(0,v)$	$(1,v)$	$\mathbb{E}_\chi[\lambda\mu] = \tilde{\lambda}\tilde{\mu}$, $\mathbb{E}_\chi[\mu] < \tilde{\mu}$
Intensive	$(1,1)$	$(1,v)$	$\mathbb{E}_\chi[\mu] = \tilde{\mu}$, $\mathbb{E}_\chi[\lambda] \lesseqgtr \tilde{\lambda}$
Both	$(1,1)$	$(0,v)$	$\mathbb{E}_\chi[\mu] = \tilde{\mu}$, $\mathbb{E}_\chi[\lambda] < \tilde{\lambda}$
	$(0,v)$	$(1,1)$	$\mathbb{E}_\chi[\lambda\mu] = \tilde{\lambda}\tilde{\mu}$, $\mathbb{E}_\chi[\mu] < \tilde{\mu}$

information frictions can also modify the nature of a social movement by affecting the intensive and extensive margins in opposite ways. If the game converges on $(1,1)$, radicals might underestimate their share and fail to coordinate on a smaller but more violent movement $(0,v)$, which they prefer for some combination of parameters. If $(0,v)$ is played repeatedly, moderates might underestimate their share and refrain from protesting, which would convince the radicals from joining a large, peaceful movement.

2.3 The effect of social media

How does political activity on social media affect the dynamics of offline protests? We modify the timeline in Figure 2 by dividing each period t into two subperiods (see Figure 3): 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 $\{0,1,v\}$ among n randomly selected individuals is revealed to all subsequent cohorts.

Figure 3: Timeline with social media



The payoffs to online participation decisions are given by the equations (1) and (2), except that the costs $\underline{c}$ and $\overline{c}$ are discounted by factors $\kappa_1 \in (0,1)$ and $\kappa_v \in (0,1)$ respectively. This reflects the fact that online participation in the movement is less individually costly than the corresponding offline action. The introduction of social media technology lowers all participation thresholds.[19] In particular, if $(0,1)$ is the online equilibrium, then it is also the offline equilibrium. Similarly, if $(1,v)$ is the offline equilibrium, then it is also the online equilibrium.

Asymptotic results. Our main insight is that social media facilitates learning about the population's propensity to protest and hence reduces the occurrence of learning traps. This happens because the cost of political participation is smaller online than offline; as a result, the equilibrium played online might differ from that played offline, and the additional information that the online equilibrium reveals might help protesters shed their wrong beliefs about λ and μ.[20] At the limit where online participation is costless ($\kappa_1 \to 0, \kappa_v \to 0$), political expression on social media reveals individuals' preferences perfectly, learning is complete, and coordination failures do not happen. Proposition 2 formalizes these observations.

Proposition 2 *For each category of learning trap identified in Proposition 1, the space of parameters $[\chi, (\tilde{\lambda}, \tilde{\mu})]$ conducive to it is strictly smaller in the version of the game with social media than in the version without it. At the limit where $\kappa_1 \to 0, \kappa_v \to 0$, no learning trap is possible.*

The effect of social media exhibits some asymmetry: when the true distribution of preferences is conducive to a large mobilization, social media can give birth to a movement that would otherwise never have started. Conversely, when discontent is minor, the population would have found this out eventually even if a movement had started in the streets based on an over-optimistic prior. In that case, social media has no effect, apart from possibly hastening the extinction of the movement.

Social media is, however, a double-edged sword in that it makes all learning traps less likely, including those in which incomplete learning is the only thing that precludes the

[19] In certain contexts, one might want to assume that $\kappa_v > 1$, reflecting that radical action is less risky offline than online due to better anonymity. In that case, social media would be ineffective at helping radicals coordinate: for instance, it would not eliminate the learning trap of a population stuck in the self-confirming equilibrium $(1,1)$ while the full-information equilibrium is $(1,v)$.

[20] There is one case where online protests do not reveal any additional information, even though the equilibria played online and offline are different: when $(0,1)$ is played offline and $(0,v)$ is played online.

rise of a violent movement. Indeed, social media helps all groups coordinate: radicals, who might stay peaceful only because they underestimate their number, can also benefit from the existence of social media as a cheaper coordination device.[21]

Short-term dynamics. In addition to the asymptotic results of Propositions 1 and 2, the model can also shed light on short-term empirical dynamics by making specific predictions about the co-evolution of beliefs and behavior. We illustrate this fact by limiting our attention to the classic situation where social media was instrumental in the launch and initial momentum of a protest movement. To that end, we enrich the model by assuming that radical protesters can also refrain from participating ($a_R = 0$) and we posit that the population starts with pessimistic beliefs: Thus, in the absence of social media, neither margin of participation is ever activated.[22] Due to the lower cost of online mobilization, social media initiates an online movement (equilibrium $(1,1)$ 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 $1b$ (equilibrium $(1,1)$), as well as some radical expression online in period $2a$ (equilibrium $(1,v)$).

We ask what possible dynamics can follow from this initial trajectory, which we call a *crowd-in sequence*. To do so, we focus on the case where $n = \infty$ in order to guarantee that the evolution of beliefs and behavior are deterministic. Proposition 3 shows that only three dynamics are possible.

Proposition 3 *Suppose that the movement initially follows a crowd-in sequence in periods $t = 1a$, $t = 1b$ and $t = 2a$ with respective equilibria $(1,1)$, $(1,1)$, and $(1,v)$. Then, from period $t = 2b$ on, the game settles on one of the following three equilibria:*

(i) massive peaceful movement $(1,1)$;

[21] Proposition 2 states that social media reduces the space of parameters conducive to a learning trap. However, it does not mean that social media always reduces the occurrence of learning traps for a given prior belief χ_0, nor that social media increases participation for every χ_0. To see this, suppose that the population's preferences are such that $(1,1)$ is the full-information equilibrium, and is played at every period of the game without social media. By reducing the cost of violent political expression, the introduction of social media might prompt the population to play $(0,v)$ instead at every period, and hence fail to revise a pessimistic prior about $\tilde{\mu}$. This might sustain $(0,v)$ as a long-run equilibrium, or even $(0,1)$ if the share of radicals is small enough. In these examples, paradoxically, social media decreases participation by indulging radical expression and preventing the coordination of moderate participants. We formulate and prove the corresponding result in Appendix A.4.

[22] The equilibrium conditions are provided in the Appendix, section A.5.

(ii) enduring revolution $(1, v)$;

(iii) crowding out of the moderates $(0, v)$.

In addition, each of these cases is the unique equilibrium for some combination of parameter values and realizations of $\tilde{\mu}$ *and of* $\tilde{\lambda}$.

These dynamics are illustrated in Figure 4. 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 what happened during the Arab Spring of 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).[23] Last, sequence 3 shows a *crowd-in-then-crowd-out* pattern where moderates finally leave the movement. We argue, and show in Section 3, that this trajectory is a good representation of the evolution of the Yellow Vest movement.[24]

Figure 4: Diverging sequences after initial crowding-in.

	$t = 1a$	$t = 1b$	$t = 2a$	$t = 2b$	$t = 3a$	$t = 3b$
Without social media		$(0, 0)$		$(0, 0)$		$(0, 0)$
With social media (sequence 1)	$(1, 1)$	$(1, 1)$	$(1, v)$	$(1, 1)$	$(1, 1)$	$(1, 1)$
With social media (sequence 2)	$(1, 1)$	$(1, 1)$	$(1, v)$	$(1, v)$	$(1, v)$	$(1, v)$
With social media (sequence 3)	$(1, 1)$	$(1, 1)$	$(1, v)$	$(0, v)$	$(0, v)$	$(0, v)$

To study the circumstances under which the *crowd-in-then-crowd-out* sequence is likely to arise, we perform the following exercise. We fix all preference parameters of the game ($\theta_M, \theta_R, v, \alpha, \beta, \gamma, \kappa_1, \kappa_v, \underline{c}$ and $\overline{c}$), as well as the prior beliefs $\chi_0(\lambda, \mu)$, and we assume that the value of $\tilde{\mu}$ is the same in all sequences. We then show in Proposition 4 that the *crowd-in-then-crowd-out* sequence prevails over the other two for the highest values of $\tilde{\lambda}$.

[23] Brummitt, Barnett and D'Souza (2015) propose a different model of revolutions during the Arab Spring based on the existence of tipping points.

[24] Sequences 2 and 3 are illustrated in the (θ_M, θ_R) plane in Appendix Figure A.1. The latter occurs because, all else being equal, θ_M is lower than in the former. Conversely, these two sequences cannot be ranked with respect to θ_R.

Proposition 4 *Suppose that the realizations $\tilde{\lambda}_1$, $\tilde{\lambda}_2$ and $\tilde{\lambda}_3$ of λ give rise respectively to sequences 1, 2 and 3 in Figure 4. Then:*

(i) $\tilde{\mu} > \mathbb{E}_{\chi_0}[\mu]$;

(ii) $\tilde{\lambda}_3\tilde{\mu} > \mathbb{E}_{\chi_0}[\lambda\mu]$ *and* $\tilde{\lambda}_3 > \mathbb{E}_{\chi_0}[\lambda \mid \tilde{\mu}]$;

(iii) $\tilde{\lambda}_3 > \tilde{\lambda}_1$ *and* $\tilde{\lambda}_3 > \tilde{\lambda}_2$.

In all three sequences, social media facilitates the mobilization initially, which happens when the population learns from online interactions that the propensity to mobilize (the realized $\tilde{\mu}$) is larger than expected (item (i)). Second, in a *crowd-in-then-crowd-out* dynamics, the population revises its beliefs upwards about the share of radicals (item (ii)): the true share of radicals $\tilde{\lambda}_3\tilde{\mu}$ is larger than both the population's prior $\mathbb{E}_{\chi_0}[\lambda\mu]$ and its interim belief $\mathbb{E}_{\chi_0}[\lambda \mid \tilde{\mu}]\tilde{\mu}$. Last, the true share of radicals is larger in a *crowd-in-then-crowd-out* dynamics than in any alternative—large peaceful movement or lasting revolution (item (iii)). In the latter comparison, it is precisely the high number of radicals that crowds out moderates' participation.[25]

2.4 Extensions

Government response. A popular policy instrument used by governments to control political protests is the shutting down of social media. This instrument is both used, or at least considered, by authoritarian regimes to restrain legitimate democratic movements, and by democratic regimes to contain violent protests.[26] Our model provides a framework for thinking about the effects of these policies. One important implication is that shutting down social media can have different effects depending on when it is implemented. To see this, consider the *crowd-in-then-crowd-out* sequence. A ban on social media implemented from $t = 1a$ on would prevent the movement from gaining momentum, resulting in the lowest participation equilibrium $(0, 0)$ in all future periods.

[25]The values of $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$ cannot be compared unequivocally: indeed, the movement might either stay peaceful in sequence 1 because radicals are not numerous enough to coordinate on action $a = v$, or because they are too numerous to do so without excluding moderates.

[26]During the 2019 protests in Iran, the Supreme National Security Council imposed a week-long Internet shutdown, during which the population could only access the national information network. 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.

However, banning social media from $t = 2a$ onward could prevent the radicalization of the movement and the subsequent crowding out of moderates, leading to equilibrium $(1, 1)$ in all future periods.

Paradoxically, shutting down social media once the movement has gained traction would then favor 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 et al. (2020) in the case of protests in several African countries between 2017 and 2019). Similarly, banning social media at the outset might paradoxically keep the population in the learning trap $(1, 1)$ rather than $(0, v)$. A strategic government concerned with containing peaceful and/or violent protests would have to consider the effects on both margins to decide on the optimal policy (and its timing).[27]

Biased reporting. Our analysis so far assumes that the information received by the population is unbiased (though not necessarily complete), in that it accurately reflects the shares of the different types of protesters. However, the algorithms used by social media platforms may bias the content shown to users (Levy, 2021). Similarly, it is conceivable that participants in or witnesses to a protest, when exposed to violent incidents, may tend to overestimate their frequency or magnitude.[28] We extend our model to allow for such a bias. We assume that both online and offline observations over-sample violent protesters: the probability with which a violent protester is sampled exceeds the unbiased probability by $b \geq 0$, while the probability with which a passive individual is sampled is lowered by b. As a result, in the equilibria $(0, v)$ and $(1, v)$, the action $a = v$ is shown with probability $\tilde{\lambda}\tilde{\mu} + b$. If one of these equilibria is played infinitely often, the long-run beliefs of a naive population overestimate the share of radicals by b.[29]

Intuition suggests that biased news strengthens the intensive margin of protests, as it leads radicals to overestimate their share. Things are more subtle, however, due to the possible crowd-out of moderates. To see this, note that, under a bias $b \geq 0$, the profile

[27] We could use similar arguments to analyze censorship and policing by the government (or self-policing by protest leaders) as in Shadmehr and Bernhardt (2015) and Ananyev, Zudenkova and Petrova (2019).

[28] To avoid an inconsistency between offline and online information in the long run, we focus on the case where the bias is asymptotically stable, which requires that both types of information be equally biased.

[29] While a bias would have no effect on a population of sophisticated learners, players unaware of the selection would make systematically incorrect inferences about the preferences of the population (see, e.g., Enke, 2020; Barron, Huck and Jehiel, 2024).

$[a = (1, v), \chi, (\tilde{\lambda}, \tilde{\mu})]$ is a self-confirming equilibrium if and only if $\text{supp}(\chi) \subseteq \{(\lambda, \mu) : \lambda\mu = \tilde{\lambda}\tilde{\mu} + b \text{ and } \mu = \tilde{\mu}\}$ and the following system holds:

$$\begin{cases} v\theta_R + (\gamma - \alpha)(\tilde{\lambda}\tilde{\mu} + b) \geq \overline{c} - \underline{c}, \\ \theta_M + \alpha(1 - \tilde{\lambda})\tilde{\mu} - \beta(\tilde{\lambda}\tilde{\mu} + b) \geq \underline{c}. \end{cases} \tag{3}$$

An increase in the bias b has opposite effects on the two equilibrium conditions in System 3: it can both reinforce radicals' willingness to engage in violent protests and make moderates more reluctant to participate. As a result, while a direct effect of biased reporting is to increase radicals' propensity for violent action, an indirect effect is to discourage moderates from participating. It might even be the case that $(1, v)$ would be an equilibrium under unbiased learning, but $(1, 1)$ results from biased learning as radicals refrain from violent action for fear of excluding moderates, who (irrationally) expect a large share of radicals within the movement. This observation qualifies the common wisdom that social media bias , which is often accused of radicalizing public debate, is necessarily a source of radicalization of protest movements.

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, data, and methods

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, but they drew fewer and fewer participants and eventually disappeared. We provide more elements of context in Appendix B and additional information on our data in Appendix C.

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, parking lots, but mostly roundabouts) and included specific descriptions of the planned events.[30] Many locations were chosen for their potential to block traffic and economic 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.[31] We complement this data with weekly national statistics from the Ministry of the Interior on the number of protesters and with the Yellow Vests' own monitoring system, called *Le Nombre Jaune*, which started in January 2019.[32]

Finally, 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 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 ones. Using a similar method, we also identified 617 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. Since Netvizz did not provide user identifiers associated with each message, we scraped Facebook a second time in January 2022 and collected additional basic user information for a subset

[30] Note that these are declarations of intent to demonstrate. However, since the map was created to coordinate roadblocks, there was little incentive to falsely declare intent to demonstrate. Contrary to what happens in autocratic regimes (Clarke and Kocak, 2020; Hassan, 2021), the French police did not preemptively try to lift the roadblocks.

[31] 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. 551 of the 1,632 Living Zones were affected (see Appendix Figure C.1).

[32] 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.

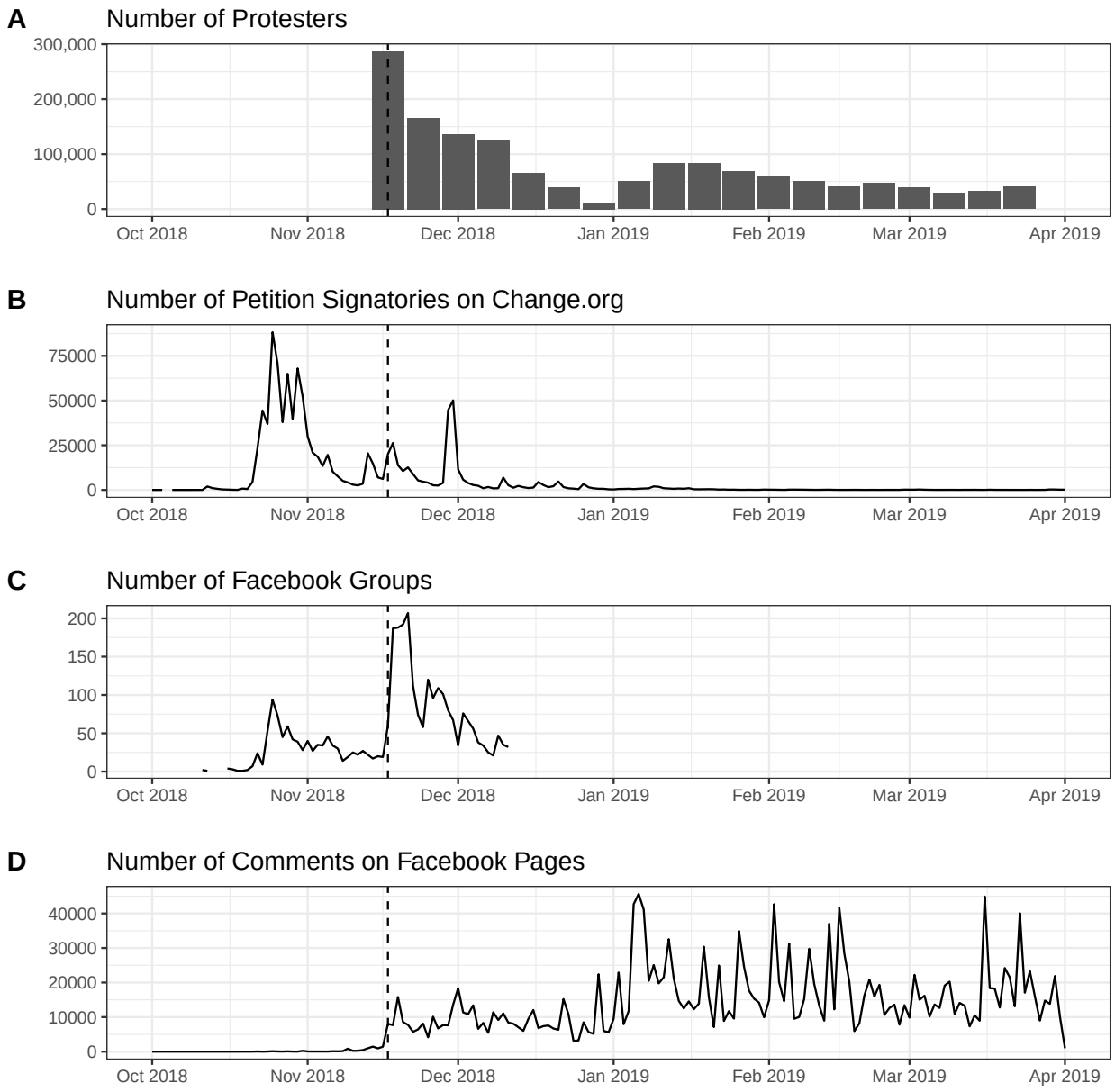
of 120,463 users.[33]

Time series. In 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 creations, and the number of comments on Facebook pages. The movement culminated in the streets during the first episode of the protests. Importantly, the decline in the number of street protesters was driven more by a decline in the size of the protests, not the number of protests (see Appendix Figure C.2). While the petition was mostly signed before 11/17, there were two distinct episodes of group creation: a first in the weeks before 11/17 and a second immediately after. This pattern suggests that 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 evolution of the intensity of discussions on dedicated Facebook pages supports this hypothesis. The discussions gained importance in January 2019 and, contrary to the weekly number of protesters, remained strongly active in the following months.

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; Ash and Hansen, 2023): a topic model, a sentiment analysis, and a political classification of the messages (see Appendix E for details). To identify the topics discussed online by the Yellow Vests, we rely on 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. To measure the emotional content of messages, we use a dictionary-based approach that assigns a sentiment score to each sentence. The sentiment score ranges between -1 and +1, where -1 corresponds to very negative sentences and +1 to very positive sentences. All topics that we classify as antagonistic are associated with more negative sentiment. Finally, to understand messages' political stance, we train a supervised learning model that pre-

[33]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 control for 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).

Figure 5: Evolution of Online and Offline Mobilizations



Notes: In Panel A, we show the number of demonstrators reported weekly by the Ministry of the Interior. In Panel B, we plot the daily number of petition signatures. In Panel C, we plot the daily number of new Facebook groups created. Finally, in Panel D, we plot the daily number of messages posted on Facebook pages. The vertical dashed line in all panels corresponds to 11/17.

dicts the party affiliation of members of the French Parliament based on their tweets. Once trained, the model predicts the probability of a given sentence being written by a specific party.

3.2 Crowding-in: the online-offline feedback loop

To document the relationship between early online and offline mobilization, and in the absence of individual-level information on both activities, we construct a dataset at the most granular level possible: the municipality. There are more than 34,000 municipalities in mainland France, whose boundaries often date back to the French Revolution. They represent the lowest level of government and many social, economic, geographical and political characteristics, listed in Appendix C.5, are available at that level.

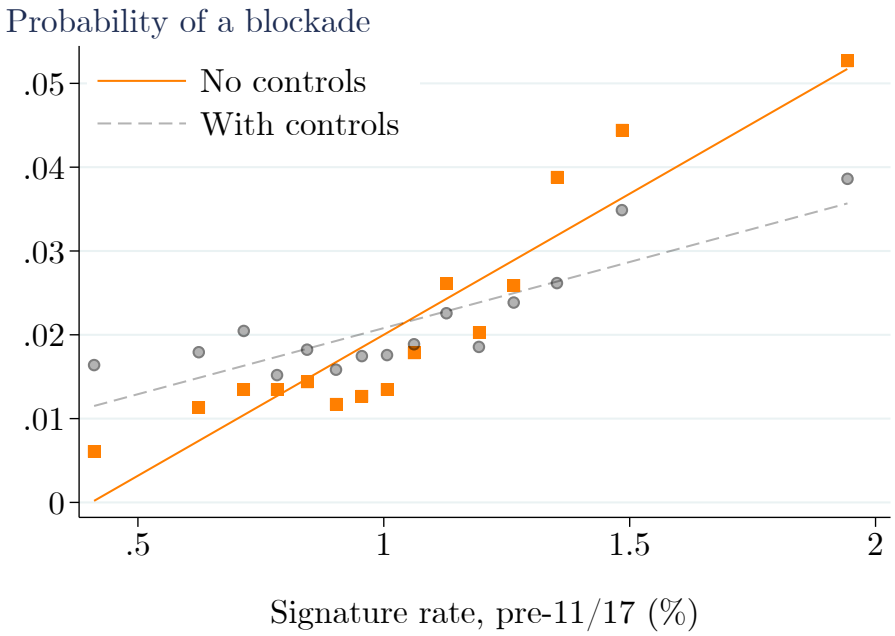
From online to offline. We start by documenting the role of early online mobilization, as measured by the number of local Facebook groups and the petition signature rate before 11/17, on the occurrence of the 11/17 protests. Figure 6 displays a positive correlation between signature rates and the probability of a roadblock. Without controls, a 1 p.p. increase in the signature rate is associated with a 3 p.p. increase in the probability of a roadblock. Controlling for local characteristics and Living Zone fixed effects attenuates this correlation, but it remains quantitatively meaningful.

The case of local Facebook groups is even more straightforward, as many of them were created with the stated purpose of organizing the 11/17 protests. Municipalities that were not blocked on 11/17 were associated with almost zero Facebook groups before 11/17, while the soon-to-be-blocked municipalities had, on average, 0.44 groups. Figure 7 shows that controlling for local characteristics and Living Zone fixed effects reduces this gap but that it remains sizable. Consistently with Qin, Strömberg and Wu (2017), this pattern confirms that the close monitoring of social media may help predict where protests are more likely to occur.[34] According to our model, members of these groups received a positive signal about the magnitude of the mobilization potential, which convinced them to participate in the 11/17 protests. Admittedly, these groups were also used to share practical information about these protests. However, this was not the case on the Change.org platform, whose primary goal was to provide information about the evolution of the number of signatories.

Since both early online activity and the 11/17 roadblocks were associated with discontent, these positive relationships may not be very informative about causality. Therefore, as a robustness check, we instrument our measures of early online activity with the presence of a 4G antenna in the municipality prior to 11/17. Access to 4G improves signal quality and thus the time people spend on their phones, which should increase the

[34] These new monitoring capabilities are not without risks, especially if online conversations allow authoritarian regimes to identify dissenters (Rød and Weidmann, 2015; Earl, Maher and Pan, 2022; Andirin, Neggers, Shadmehr and Shapiro, 2022).

Figure 6: Early petition signatures and the probability of a roadblock

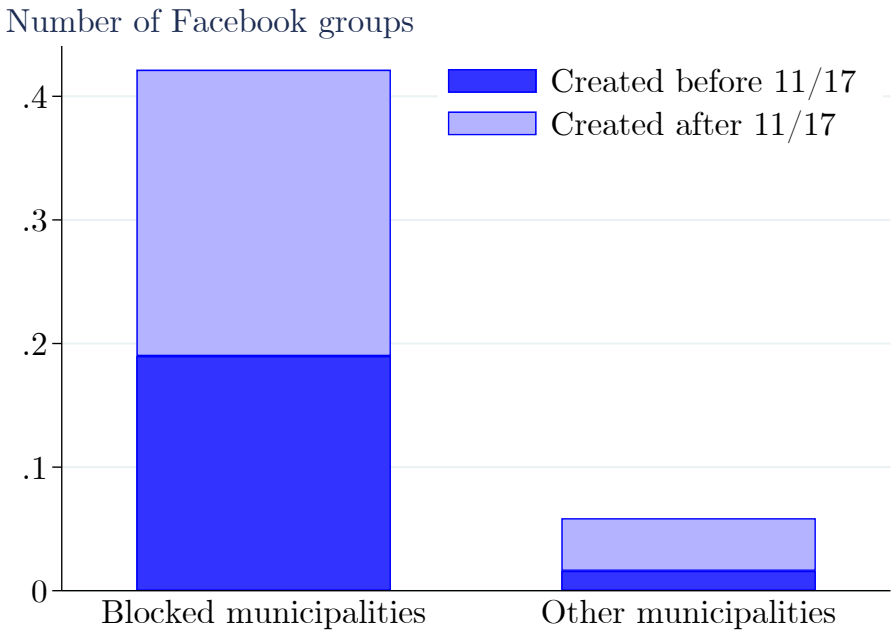


Notes: Binscatter plot of the relationship between the signature rate per capita before 11/17 and the unconditional or conditional probability of a roadblock in the municipality. The list of controls is detailed in Appendix C.5.

likelihood that they will hear about the petition or coordinate to form a local Facebook group. The roll-out of 4G in France was about half complete at the time. 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. The results are presented and discussed in Appendix D.1 and they confirm that early online activity had a large direct positive impact on the probability of organizing a roadblock, consistently with an extensive body of research.

From offline back to online. We then turn to the opposite direction of this link and ask a less-studied question: do street protests encourage further online mobilization? Indeed, the initial Yellow Vests street protests were large, mostly peaceful, and showed that a large part of the population was sympathetic to the movement and inclined to participate. According to our model, this may have further increased optimism about the share of non-passives in the population, and increased the intensity of subsequent

Figure 7: The rebound effect: Local Facebook groups and the 11/17 roadblocks



Notes: Average number of local Facebook groups in municipalities that experienced a roadblock on 11/17 and in other municipalities, net of local characteristics and Living Zone fixed effects. The list of controls is detailed in Appendix C.5. For groups created after 11/17, we also control for the number of groups created before 11/17 and for the petition signature rate before 11/17.

online mobilization.[35] As shown in Panel C of Figure 5, such a rebound effect occurred in the aggregate: many Facebook groups were created in the immediate aftermath of the 11/17 protests. While this pattern may be coincidental, Figure 7 shows that the effect was concentrated in municipalities that had experienced a roadblock on 11/17. On average, they welcomed one new Facebook group in the four weeks following 11/17, which, after controlling for local characteristics and Living Zone fixed effects, corresponds to over a doubling of their net stock of groups. Conversely, the other municipalities experienced only very little group creation.

Once again, such a positive correlation may not be informative about causality. As stated above, the positive correlation between discontent and both roadblocks and post-11/17 online mobilization would induce an upward bias. Conversely, intertemporal substitutability between the different stages of the protests could induce a downward bias, for example if the average willingness to protest decreases over time. Therefore,

[35]Consistently with this information channel, we provide suggestive evidence that weekly street protests were associated with a sharp increase in Google queries about the Yellow Vests on Facebook (see Appendix Figure C.6).

we propose an instrumental variable strategy based on the spatial dispersion of roundabouts in French municipalities. Roundabouts are attractive protest locations because they enable demonstrators to block several roads simultaneously and are easy to camp on. At the same time, they are widely recognized as architectural fads, and, most of the time, they can be replaced with traffic lights. All else equal, roundabouts are easy-to-block locations that lower the cost of organizing a blockade independently of the local demand for protest. Results are presented and discussed in Appendix D.2 and they confirm that street protests did trigger additional online mobilization, both on Facebook and on Change.org. According to our 2SLS estimates, a roadblock in a municipality increased the number of new local Facebook groups by 1.2, which is slightly higher than the reduced-form figure reported in Figure 7.

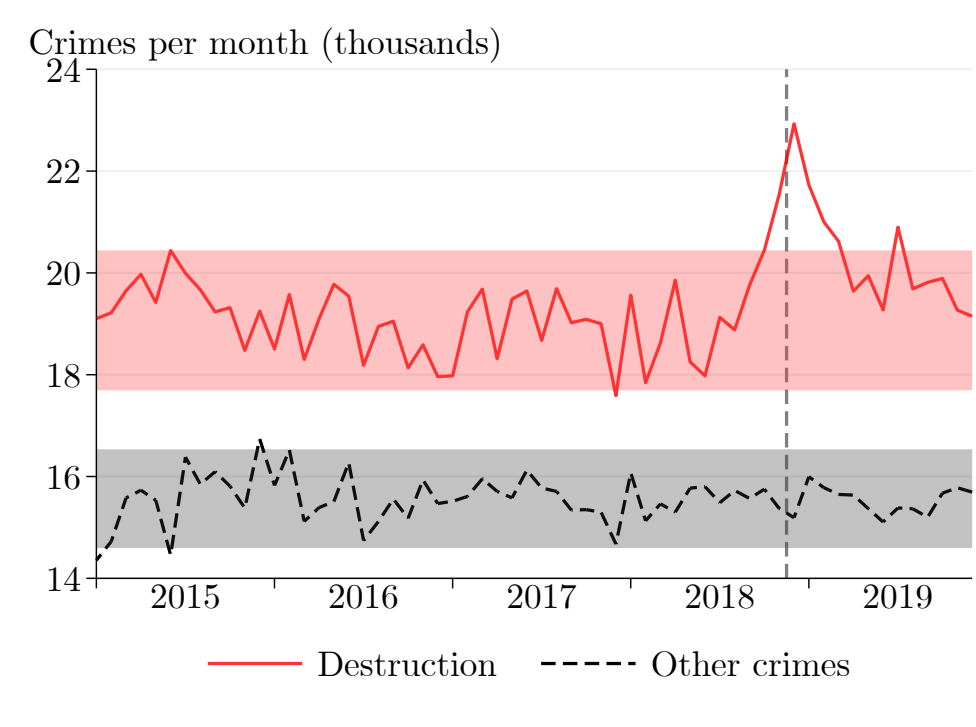
Protest persistence. To close the online-offline feedback loop, we extend the analysis to later demonstrations. According to the Yellow Vests' own monitoring system (*Le Nombre Jaune*), only 55% (405 out of 741) of the municipalities that we observe as blocked on 11/17 also experienced protests between January and May 2019.[36] Appendix Figure D.3 shows that the offline-online rebound effect displayed in Figure 7 was fully concentrated in the municipalities that experienced further protests in 2019. Conversely, the municipalities blocked on 11/17 but where no protest took place in 2019 experienced no further group creation after 11/17. By mid-December 2018, the municipalities that only joined the movement after 11/17 had, on average, more local Facebook groups than those early dropouts. In other words, street mobilization declined in places where online mobilization had also declined.

3.3 Violence and the crowding-out of moderates

To further understand the decline in mobilization, we turn to the intensive margin of protests: their intensity. According to our theoretical framework, violent protests can crowd out moderate protesters and reduce the size of subsequent mobilizations. We proceed in two steps: first, we use our municipal dataset to show descriptively that local street violence was associated with the subsequent formation of smaller online communities and smaller street protests. We then use textual analysis to show that new

[36] As explained above, the movement has also changed over time, with an increasing concentration of weekly protests in the larger cities. We cannot therefore rule out the possibility that some of the participants in the initial roadblocks may have moved on to protest in nearby towns.

Figure 8: Street violence



Notes: Monthly time series of destruction-related offenses and other offenses (related to vehicle theft and drug trafficking) in continental France, after accounting for month fixed effects. The shaded areas correspond to the 95% tolerance intervals based on observations prior to November 2018. The dashed line corresponds to 11/17.

online communities were smaller because radicalized discussants drove the moderates away.

Street violence and the shrinking of later protests. Although the 11/17 roadblocks were largely peaceful, the movement became quite violent in the following weeks. To document street violence, we first use monthly data from the Ministry of the Interior, which gives the number of offenses recorded by the police. We isolate one class of offenses: "destruction of public and private property," which we use as a proxy for rioting. Figure 8 shows that destruction peaked at the turn of 2019, reaching a level equal to 5 standard deviations above the pre-protest average in December 2018.

We then use the annual geolocated version of this data to create an index of street violence at the municipal level.[37] Conditional on our set of local controls and Living Zone

[37]See Appendix C.8 for details. While blocked municipalities experienced significantly more destruction than other municipalities in 2018, this was not the case for other recorded crimes. Conversely, both types of municipalities experienced similar levels of

fixed effects, we estimate that groups formed before 11/17 were 11% larger (measured by number of members) in blocked municipalities with above-median levels of violence in 2018. Conversely, groups formed after 11/17 in these municipalities were 8% smaller than in other blocked municipalities, even after controlling for the average size of groups formed before 11/17.[38] We then replicate this analysis for street mobilizations in 2019, for which we have information on the number of protests and their size. We construct a measure of the size of subsequent street protests as the ratio of the median protest size (conditional on our set of local controls and Living Zone fixed effects) to the municipal population. On average, among municipalities with persistent mobilization, those with above-median levels of violence had a relative protest size of 5%, compared to 7% for municipalities below the median.[39] While these differences cannot be interpreted as causal, they suggest that a more violent local mobilization was associated with a shrinkage of later protests. According to our model, this shrinking pattern suggests that high online mobilization prior to 11/17 led radical protesters to turn to violent action during the 11/17 protests, but that moderates were deterred by these violent protests and did not participate in subsequent online and offline mobilizations.

The two margins of online radicalization. To assess whether this shrinking pattern was driven by the departure of moderates, and in the absence of panel data on street protesters, we turn to another source of information that allows us to measure individual protesting activity and follow protesters over time: the discussions that took place on the Yellow Vests' Facebook pages. 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 increased by 15 p.p. over the period. Similarly, the share of messages classified as having negative sentiment increased by 8 p.p.[40] Of course, some messages that contain antagonistic elements or show negative sentiments may also reflect the fact that online discussants are describing vi-

crime in 2017 and 2019.

[38]The difference between the relative effect of violence on the size of groups before 11/17 and the relative effect of violence on the size of groups after 11/17 is statistically significant at the 5% confidence level.

[39]This difference is statistically significant at the 5% level. Using the average or maximum number of protesters instead of the median yields similar results (9% vs. 7% and 29% vs. 21%).

[40]While negative sentiment could encompass very different emotions, we provide suggestive evidence that anger drove this increasing pattern (see Appendix Figure E.4).

olent 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, and we observe that the share of messages with far-left or far-right language increased by 6 p.p., suggesting a polarization of online discussions.[41]

In summary, online discussions have become more antagonistic, negative, and polarized: for lack of a better word and to stay close to the terminology used in Section 2, we refer to these combined characteristics under the umbrella term *radicalization*.[42] According to our theoretical framework, this radicalization could have been driven by two different margins: First, moderate users may have gradually left the movement or been replaced by more radical users. We refer to this attrition effect as the "extensive margin" of radicalization. Alternatively, active users may have become more radical over time. We refer to such individual changes as the "intensive margin" of radicalization.

Anecdotally, we can observe these two margins as the tension between moderates and radicals unfolds on Facebook pages. Appendix Table E.3 shows examples of messages sent by online protesters. Some protesters justify and support street violence. For example, a protester writes: *"Even today, we are obliged to call on our traditions of violence to defend our right to a decent life."* Others condemn it and worry it will discredit the movement. For instance, a protester analyzes: *"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."* In line with the role played by the extensive margin, some discussants are debating whether to participate 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 your support. You are courageous."* Yet, in December and January, his or her tone markedly changes with messages such as: *"Reduce these ******** to nothing."* and *"All corrupt, these ********."* This anecdotal evidence suggests both margins are potentially meaningful.

To quantify their relative contributions, we exploit the panel dimension of the data and the fact that we can follow individual (de-identified) discussants over time. To

[41] This finding is consistent with polling data showing that the decline in popular support for the movement was mostly driven by centrist voters (see Appendix Figure C.5).

[42] This concept is quite equivocal. It is often used to qualify identity-based politics and ideology (see, e.g., Carvalho and Sacks, 2023), which does not apply here.

isolate the intensive margin of radicalization, we can assess whether the average user has become increasingly likely to post radical messages. To isolate the extensive margin, we can assess whether the pool of active users becomes increasingly populated with users who (on average) post more radical messages. We estimate the following equation:

$$Y_{s,i,t} = \delta_i + \gamma_t + \varepsilon_s, \tag{4}$$

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.

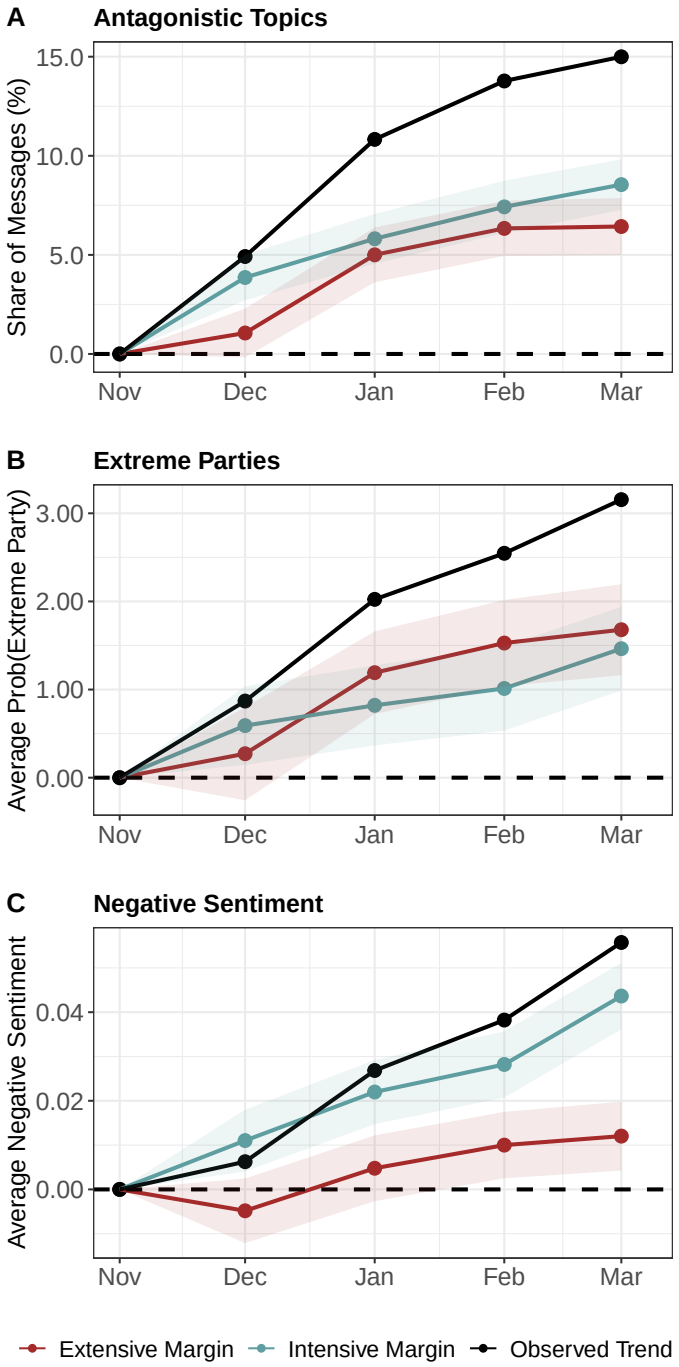
We can then leverage estimates of user and time fixed effects to decompose the rise of online radicalism into an intensive and extensive margin. Indeed, the average level of radical sentences during month t, $\mathbb{E}_t[Y]$, can be expressed as:

$$\mathbb{E}_t[Y] = \underbrace{\mathbb{E}_t[\delta]}_{\text{Extensive margin}} + \underbrace{\gamma_t}_{\text{Intensive margin}}, \tag{5}$$

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 expression 5 corresponds to the average propensity to post radical sentences for users active during the 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 expression 5 corresponds to an increase in the propensity of any given user to post a radical sentence at a given time.

Figure 9 presents a decomposition of our radicalization measures using the empirical counterpart of Equation 5. In Panel A, the outcome variable is a dummy variable indicating whether a message was associated with an antagonistic topic. In Panel B, the outcome variable is the probability that a message was associated with an extreme political party. In Panel C, the outcome variable is the negative sentiment score associated with a sentence, which takes values between -1 (very positive) and 1 (very negative). For all three dependent variables, our decomposition suggests that both margins contributed to the radicalization of Facebook content. Quantitatively, both margins played a similar role in two of the three measures. Moreover, the effect of the extensive margin appears to be slightly delayed relative to the intensive margin, suggesting that the radicalization of some discussants triggered the defection of the more moderate ones.

Figure 9: Extensive and Intensive Margins of Radicalization



Notes: This figure decomposes the increase in online radicalism using Equation 5. Panel A presents estimates for the probability of posting a sentence associated with an antagonistic topic. Panel B presents estimates for the average probability of writing a sentence associated with a politically extreme party (i.e., on the far left or the far right). Panel C presents estimates for negative sentiment. We compute standard errors via the nonparametric bootstrap with 1000 iterations and plot confidence intervals at the 95% confidence level.

Moderates leave radicalized discussions. To better understand what drives the crowding out of moderates, we use the previous framework to measure the impact of discussion radicalization on the online mobilization of different types of protesters. For each measure of radicalism, we first estimate discussant fixed effects using Equation 4. Then, on the sample of discussant-by-page-by-month observations, we estimate the following equation:

$$\mathbb{P}(\text{Exit})_{i,p,t} = \sum_q \beta_q \left(\mathbb{1}_{\delta_i \in q} \times \mathbb{E}_{p,t}\left[\delta\right] \right) + \zeta_i + \xi_{p,t} + \boldsymbol{X}_{i,p,t}\eta + \varepsilon_{i,p,t}, \tag{6}$$

where $\mathbb{P}(\text{Exit})_{i,p,t}$ is the probability that discussant i stops posting on page p after month t,[43] $\mathbb{1}_{\delta_i \in q}$ 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.[44] For the estimation, we replace expectations and quantiles of δ_i by their empirical counterparts (using our estimates of Equation 4).

Our results are summarized in Figure 10, which breaks down individual radicalism into quintiles.[45] 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 her probability to stop posting on that page by 4 to 9 p.p., or 6 to 14% of the baseline probability. 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.

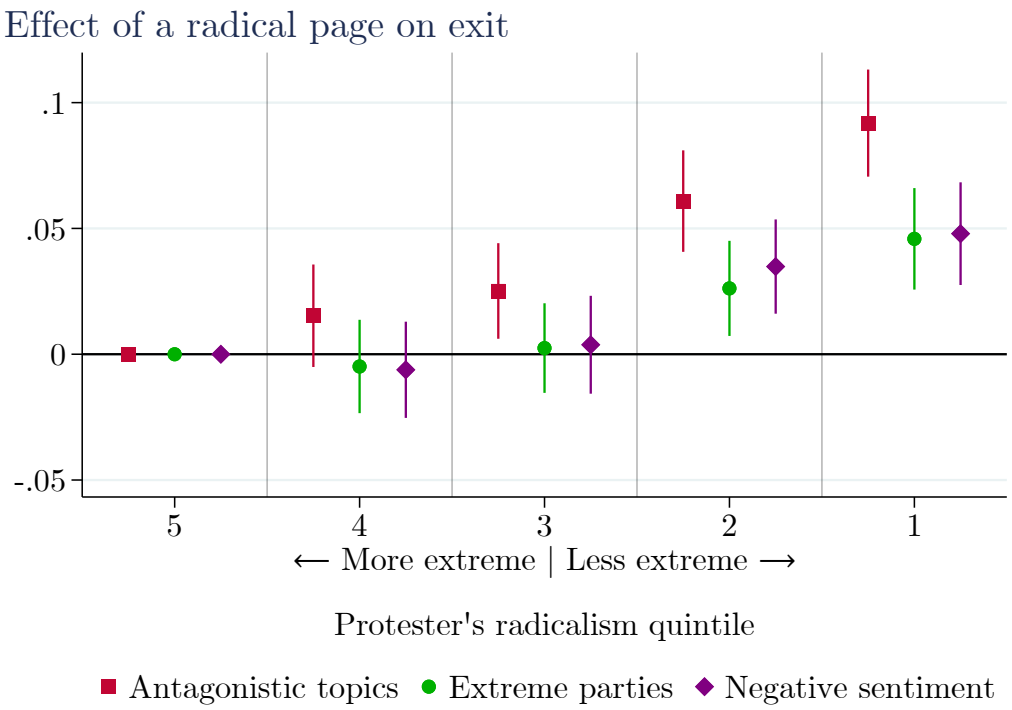
We evaluate the robustness of this result along several dimensions. First, one may consider that a better measure of page radicalism would be the radicalism of the average posted sentence ($\mathbb{E}_{p,t}[Y]$), rather than the average value of discussants' radicalism fixed

[43] Hence, for this second stage, we restrict the estimation sample to pages that are still active the following months. In practice, this mostly means dropping March 2019 from the sample.

[44] In practice, we control for the number of sentences posted by the discussant during month t, either on page p or on other pages. The former is negatively correlated with the exit probability, and the latter is positively correlated. We also control for a binary variable indicating whether the discussant had already posted on the page before month t.

[45] In robustness tables, we use a binary variable indicating whether the discussant's fixed effect is below or above the median instead of quintiles, for brevity.

Figure 10: Crowding-out over the distribution of protesters' radicalism



Notes: This figure shows the OLS estimates of β_1, β_2, β_3 and β_4 of Equation 6 and their associated 95% confidence intervals, with standard errors clustered at the discussant level. β_5 is set to zero by normalization. Each series refers to a different measure of radicalism used as outcome in Equation 4 and then used as a right-hand-side variable in Equation 6: "Antagonistic topics" refers to the probability of posting a sentence associated with an antagonistic topic, "Extreme parties" refers to the probability that the sentence is associated with a politically extreme party and "Negative sentiment" is equal to (-1) $\times$ the sentiment score associated with the sentence.

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 Appendix Figure E.7, the results are remarkably similar if we use this alternative measure of page radicalism.[46]

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,

[46]Similarly, the results are robust to computing page radicalism without including the sentences posted by the discussant herself – See Appendix Figure E.8.

which suggests that moderate discussants sort across pages according to their tolerance for radical discussion, even if they do not post radical messages themselves.[47]

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 Figure 9 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 Tables E.7 and E.8, 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.

Finally, to check whether the crowd-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 Appendix Figure E.9 confirm that crowd-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 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 Appendix Figure E.10).

Alternative sources of radicalization. While these pieces of evidence are 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 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. However, the precise chronology of the movement suggests that the street

[47] Appendix Table E.8 shows that this increasing pattern is not driven by the sample selection that results from these more stringent specifications.

protests turned violent very quickly (in the very days after 11/17), before the official policy announcements and despite the initial roadblocks being largely met with police restraint. In addition, this mechanism cannot explain the crowding out of online protesters at the Facebook page level – since all protesters were exposed to the government response.

More directly related to our framework, social media may also be particularly conducive to radicalization, because of how platforms organize discussions. In Appendix E.6 we describe a strategy to identify the effect of Facebook's recommendation algorithm on the visibility of radical statements. We use the structure of online discussions, where comments are not displayed in chronological order but instead reordered by the platform, and we show that, consistently with a potential algorithmic bias, discussants on the Yellow Vest Facebook pages were over-exposed to radical content. However, as argued in section 2.4, such a bias by itself would not necessarily lead to the radicalization of online discussions and to more violence on the streets.

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 since the twentieth century. To do so, it draws on a salient feature of contemporary protest movements: their use of social media. We propose a simple 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 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 et CEVIPOF 2019.

Alsulami, Amer, Anton Glukhov, Maxim Shishlenin, and Sergei Petrovskii, "Dynamical modelling of street protests using the Yellow Vest Movement and Khabarovsk as case studies," *Scientific Reports*, 2022, *12* (1), 20447.

Ananyev, Maxim, Galina Zudenkova, and Maria Petrova, "Information and communication technologies, protests, and censorship," 2019. Working paper.

Andirin, Veli, Yusuf Neggers, Mehdi Shadmehr, and Jesse M Shapiro, "Real-time Surveillance of Repression: Theory and Implementation," Working Paper 30167, National Bureau of Economic Research June 2022.

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.

Arguedas, A Ross, 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.

Aridor, Guy, Rafael Jiménez-Durán, Ro'ee Levy, and Lena Song, "The Economics of Social Media," *Journal of Economic Literature*, Forthcoming.

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.

Barron, Kai, Steffen Huck, and Philippe Jehiel, "Everyday Econometricians: Selection Neglect and Overoptimism When Learning from Others," *American Economic Journal: Microeconomics*, August 2024, *16* (3), 162â98.

Bastos, Marco T., Dan Mercea, and Arthur Charpentier, "Tents, Tweets, and Events: The Interplay Between Ongoing Protests and Social Media," *Journal of Communication*, 2015, *65* (2), 320–350.

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," *The 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.

Carvalho, Jean-Paul and Michael Sacks, "Radicalisation," *The Economic Journal*, 10 2023, *134* (659), 1019–1068.

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," 2022. Working paper.

Della Porta, Donatella and Mario Diani, *Social Movements: An Introduction*, Wiley-Blackwell, 2020.

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.

Earl, Jennifer, Thomas V. Maher, and Jennifer Pan, "The Digital Repression of Social Movements, Protest, and Activism: A Synthetic Review," *Science Advances*, 2022, *8* (10), eabl8198.

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.

Enke, Benjamin, "What you see is all there is," *Quarterly Journal of Economics*, 2020, *135* (3), 1363–1398.

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

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, GermÃ¡n 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.

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.

Ives, Brandon and Jacob S. Lewis, "From Rallies to Riots: Why Some Protests Become Violent," *Journal of Conflict Resolution*, 2020, *64* (5), 958–986.

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.

Le Galès, Patrick, "The Rise of Local Politics: A Global Review," *Annual Review of Political Science*, 2021, *24* (1), 345–363.

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," *The 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," *The 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.

Madestam, Andreas, Daniel Shoag, Stan Veuger, and David Yanagizawa-Drott, "Do Political Protests Matter? Evidence From the Tea Party Movement," *Quarterly Journal of Economics*, 2013, *128* (4), 1633–1685.

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.

Pariser, Eli, *The Filter Bubble: What the Internet is Hiding from You*, Penguin UK, 2011.

Qin, Bei, David Strömberg, and Yanhui Wu, "Why Does China Allow Freer Social Media? Protests Versus Surveillance and Propaganda," *Journal of Economic Perspectives*, 2017, *31* (1), 117–40.

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.

Rød, Espen Geelmuyden and Nils B Weidmann, "Empowering Activists or Autocrats? The Internet in Authoritarian Regimes," *Journal of Peace Research*, 2015, *52* (3), 338–351.

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," *The 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

Contents

A Proofs

A.1 Analysis of the Stage Game

We start by collecting the conditions for all four possible equilibria, for given beliefs $\mathbb{E}[\mu]$ and $\mathbb{E}[\lambda\mu]$.

The profile $(0,1)$ is an equilibrium if and only if

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

The profile $(1,1)$ is an equilibrium if and only if

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

The profile $(0,v)$ is an equilibrium if and only if

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

The profile $(1,v)$ is an equilibrium if and only if

$$\begin{cases} \theta_M + \alpha\mathbb{E}[(1-\lambda)\mu] - \beta\mathbb{E}[\lambda\mu] \geq \underline{c}, \\ v\theta_R + (\beta+\gamma)\mathbb{E}[\lambda\mu] \geq \overline{c} - \underline{c}. \end{cases}$$

We then examine all regions of Figure 1 in turn, defining the thresholds as in the text.

First case: $\theta_R > \overline{\theta}_R$ This implies that $a_R = v$ is a dominant strategy for radicals. If $\theta_M < \overline{\theta}_M$, then $(0,v)$ is the unique equilibrium. If instead $\theta_M \geq \overline{\theta}_M$, then $(1,v)$ constitutes an equilibrium; there is a region where it coexists with $(0,v)$, but moderates always prefer $(1,v)$ to $(0,v)$ when $(1,v)$ is an equilibrium. This implies the characterization

of Figure 1 for $\theta_R > \overline{\theta}_R$, where the equilibrium $(1, v)$ is played on the vertical line at $\theta_M = \overline{\theta}_M$.

Second case: $\underline{\theta}_R < \theta_R \leq \overline{\theta}_R$ Note that the condition $\theta_R > \underline{\theta}_R$ implies that, when $(0, 1)$ and $(0, v)$ coexist, radicals prefer coordinating on $(0, v)$; similarly, when $(1, 1)$ and $(1, v)$ coexist, radicals prefer coordinating on $(1, v)$.

Suppose first that $\theta_M < \underline{\theta}_M$. This implies that moderates play $a_M = 0$, and $(0, 1)$ and $(0, v)$ are the two possible equilibria, the latter being preferred by radicals.

Suppose now that $\underline{\theta}_M \leq \theta_M < \overline{\theta}_M$. In that region: (i) $(1, 1)$ and $(0, v)$ are equilibria, the latter because $\theta > \underline{\theta}_R \Rightarrow v\theta_R + (\beta + \gamma)\mathbb{E}[\lambda\mu] \geq \bar{c} - \underline{c}$; (ii) $(1, v)$ is not an equilibrium since $\theta_M < \overline{\theta}_M$; (ii) $(0, 1)$ is an equilibrium in a sub-region but it is ranked by moderates strictly below $(1, 1)$.

The only question is thus whether our selection criteria allow deciding between $(1, 1)$ and $(0, v)$. Moderates prefer the former strictly if $\theta_M > \underline{\theta}_M$, and are indifferent if $\theta_M = \underline{\theta}_M$.

Radicals prefer $(1, 1)$ as well in a strict sense if and only if $\theta_R < \theta_R^*$, and are indifferent if $\theta_R = \theta_R^*$. We therefore have two cases to consider.

If $\theta_R^* > \overline{\theta}_R$, i.e. if $\gamma\mathbb{E}[\lambda\mu] < \alpha\mathbb{E}[\mu]$, the fact that $\theta_R \leq \overline{\theta}_R < \theta_R^*$ implies that radicals prefer $(1, 1)$ strictly, and therefore the Pareto criterion selects $(1, 1)$ on the entire region $[\underline{\theta}_M, \overline{\theta}_M) \times (\underline{\theta}_R, \overline{\theta}_R]$.

Otherwise, we have an additional threshold $\theta_R^* \leq \overline{\theta}_R$ such that: $(1, 1)$ is played when $\theta_R \leq \theta_R^*$ and $\theta_M > \underline{\theta}_M$; $(0, v)$ is played if $\theta_R > \theta_R^*$ and $\theta_M = \underline{\theta}_M$ as radicals strictly prefer it and moderates are indifferent; finally, we leave the equilibrium indeterminate if $\theta > \theta_R^*$ and $\theta_M > \underline{\theta}_M$, as moderates then strictly prefer $(1, 1)$ but radicals strictly prefer $(0, v)$.

Suppose finally that $\overline{\theta}_M \leq \theta_M$. Then $(1, v)$ is an equilibrium. In addition, every other possible equilibrium is dominated: $(0, v)$ since moderates prefer $(1, v)$ (at least weakly) and radicals prefer $(1, v)$ strictly; $(1, 1)$ since radicals strictly prefer $(1, v)$, and $(0, 1)$ since moderates weakly prefer $(1, v)$ and radicals strictly prefer $(1, v)$.

This yields the characterization of Figure 1 for $\underline{\theta}_R < \theta_R \leq \overline{\theta}_R$.

Third case: $\theta_R \leq \underline{\theta}_R$ Suppose first that $\theta_M < \underline{\theta}_M$. On that region, $(0, 1)$ is an equilibrium; there exists a subregion where $(0, v)$ is also an equilibrium, but $(0, 1)$ is preferred by the radicals if $\theta_R < \underline{\theta}_R$, and if $\theta = \underline{\theta}_R$ we also break ties in favor of $(0, 1)$, which involves less participation.

Suppose now that $\underline{\theta}_M \leq \theta_M$. Then $(1, 1)$ constitutes an equilibrium. In addition, $(1, 1)$ Pareto-dominates every alternative equilibrium: $(0, v)$ because$\theta_R \leq \underline{\theta}_R < \theta_R^*$, $(1, v)$

because radicals are at most indifferent and moderates strictly prefer $(1,1)$, and $(0,1)$ because moderates are at most indifferent (if $\theta_M = \underline{\theta}_M$) and because radicals strictly prefer $(1,1)$. This yields the characterization of Figure 1 for $\theta_R \leq \underline{\theta}_R$. ∎

A.2 Proof of Proposition 1

In the following we fix a learning trap $(a, \chi, (\tilde{\lambda}, \tilde{\mu}))$, and distinguish cases as a function of the equilibrium a played.

First case: $a = (1, v)$ The distribution $f(. \mid (1, v), \tilde{\lambda}, \tilde{\mu})$ 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: $a = (1, 1)$ The distribution $f(. \mid (1, 1), \tilde{\lambda}, \tilde{\mu})$ identifies $\tilde{\mu}$, and thus $\mathbb{E}_\chi[\mu] = \tilde{\mu}$. The fact that $a = a^*(\chi) = (1, 1)$ then implies that $\theta_M + \alpha\tilde{\mu} > \underline{c}$, which in turn implies that $a^*(\delta_{\tilde{\lambda},\tilde{\mu}}) \in \{(0, v), (1, v)\}$. That $a = a^*(\chi) = (1, 1)$ also implies $v\theta_R < \overline{c} - \underline{c}$.

Suppose first that $a^*(\delta_{\tilde{\lambda},\tilde{\mu}}) = (0, v)$. Since $v\theta_R < \overline{c} - \underline{c}$, this is possible only if $\gamma\mathbb{E}_\chi[\lambda] > \alpha$ so that the striped area on Figure 1 is non-trivial. Then, $a^*(\chi) = (1, 1)$ and $a^*(\delta_{\tilde{\lambda},\tilde{\mu}}) = (0, v)$ is equivalent to the following system:

$$\begin{cases} v\theta_R + \tilde{\mu}(\gamma\mathbb{E}_\chi[\lambda] - \alpha) < \overline{c} - \underline{c} < v\theta_R + \tilde{\mu}(\gamma\tilde{\lambda} - \alpha), \\ \theta_M + \alpha\tilde{\mu} > \underline{c}, \\ \theta_M + \tilde{\mu}(\alpha\mathbb{E}_\chi[1 - \lambda] - \beta\mathbb{E}_\chi[\lambda]) < \underline{c}. \end{cases}$$

The first line implies in particular that $\tilde{\lambda} > \mathbb{E}_\chi[\lambda]$. This characterizes the learning trap on the fourth row of Table 1.

Suppose now that $a^*(\delta_{\tilde{\lambda},\tilde{\mu}}) = (1, v)$, i.e. that

$$v\theta_R + (\gamma - \alpha)\tilde{\lambda}\tilde{\mu} > \overline{c} - \underline{c} \text{ and } \theta_M + \tilde{\mu}[\alpha(1 - \tilde{\lambda}) - \beta\tilde{\lambda}] > \underline{c}.$$

There are two possibilities for the fact that $a = (1, 1)$. The first (area below $(1, v)$ on Figure 1) is that

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

which implies that $\mathbb{E}_\chi[\lambda] < \tilde{\lambda}$. In that situation, radicals underestimate their number and fail to coordinate on a violent protest, which would be an equilibrium.

The second possibility (area to the left of $(1, v)$ on Figure 1) is that

$$\begin{cases} v\theta_R + \tilde{\mu}(\gamma \mathbb{E}_\chi[\lambda] - \alpha) < \bar{c} - \underline{c} < v\theta_R + (\gamma - \alpha)\tilde{\lambda}\tilde{\mu}, \\ \theta_M + \tilde{\mu}(\alpha \mathbb{E}_\chi[1-\lambda] - \beta \mathbb{E}_\chi[\lambda]) < \underline{c} < \theta_M + [\alpha(1-\tilde{\lambda}) - \beta\tilde{\lambda}] \end{cases}$$

The second row of this system implies that $\tilde{\lambda} < \mathbb{E}_\chi[\lambda]$. In that case, radicals overestimate their number and fail to coordinate on a violent protest as they (mistakenly) believe that moderates would then leave the movement. This covers the third row of Table 1.

Third case: $a = (0, v)$ The distribution $f(. \mid (0, v), \tilde{\lambda}, \tilde{\mu})$ identifies $\tilde{\lambda}\tilde{\mu}$, and thus χ satisfies $\mathbb{E}_\chi[\lambda\mu] = \tilde{\lambda}\tilde{\mu}$. Since $a = a^*(\chi) = (0, v)$, we have $v\theta_R + (\gamma - \alpha)\tilde{\lambda}\tilde{\mu} > \bar{c} - \underline{c}$.

First, if $v\theta_R > \bar{c} - \underline{c}$, there is a learning trap with $a^*(\delta_{\tilde{\lambda},\tilde{\mu}}) = (1, v)$ if and only if

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

which requires $\tilde{\mu} > \mathbb{E}_\chi[\mu]$.

Second, if $v\theta_R < \bar{c} - \underline{c}$, there is a learning trap where $a^*(\delta_{\tilde{\lambda},\tilde{\mu}}) = (1, 1)$ if and only if one of these conditions is satisfied:

$$\max\{\theta_M + \alpha \mathbb{E}_\chi[\mu], \theta_M + \tilde{\mu}[\alpha(1-\tilde{\lambda}) - \beta\tilde{\lambda}]\} < \underline{c} < \theta_M + \alpha\tilde{\mu},$$

or

$$\begin{cases} v\theta_R + \gamma\tilde{\lambda}\tilde{\mu} - \alpha\tilde{\mu} < \bar{c} - \underline{c} < v\theta_R + \gamma\tilde{\lambda}\tilde{\mu} - \alpha \mathbb{E}_\chi[\mu], \\ \theta_M + \alpha \max\{\tilde{\mu}, \mathbb{E}_\chi[\mu]\} - (\alpha + \beta)\tilde{\lambda}\tilde{\mu} < \underline{c} < \theta_M + \alpha \min\{\tilde{\mu}, \mathbb{E}_\chi[\mu]\}. \end{cases}$$

Both cases require $\mathbb{E}_\chi[\mu] < \tilde{\mu}$.

Finally, there is a learning trap $a^*(\delta_{\tilde{\lambda},\tilde{\mu}}) = (1, v)$ if and only if

$$\theta_M + \alpha \mathbb{E}_\chi[\mu] < \underline{c} < \theta_M + \tilde{\mu}[\alpha(1-\tilde{\lambda}) - \beta\tilde{\lambda}].$$

Both systems require $\tilde{\mu} > \mathbb{E}_\chi[\mu]$. These cases cover the second and fifth row of Table 1.

Fourth case: $a = (0, 1)$ The distribution $f(. \mid (0, 1), \tilde{\lambda}, \tilde{\mu})$ identifies $\tilde{\lambda}\tilde{\mu}$, and thus χ satisfies $\mathbb{E}_\chi[\lambda\mu] = \tilde{\lambda}\tilde{\mu}$. This implies that $v\theta_R + (\gamma - \alpha)\tilde{\lambda}\tilde{\mu} < \underline{c}$, and the only possible learning trap is one where $a^*(\delta_{\tilde{\lambda},\tilde{\mu}}) = (1, 1)$, which arises if and only if

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

This implies $\mathbb{E}_\chi[\mu] < \tilde{\mu}$. This case covers the first row of Table 1. ∎

A.3 Proof of Proposition 2

We illustrate the logic using the first row of Table 1 as an example and skip the proof for the other cases, as it relies on a similar argument. Recall from the previous section that this learning trap arises in the absence of social media if and only if the following system holds:

$$\begin{cases} v\theta_R + (\gamma - \alpha)\tilde{\lambda}\tilde{\mu} < \overline{c} - \underline{c}, \\ \theta_M + \alpha\mathbb{E}_\chi[\mu] < \underline{c} < \theta_M + \alpha\tilde{\mu}, \\ \text{supp}(\chi) \subseteq \{(\lambda, \mu) : \lambda\mu = \tilde{\lambda}\tilde{\mu}\}. \end{cases} \tag{7}$$

This learning trap survives to the introduction of social media if and only if the online equilibrium played under belief χ is $(0,1)$ or $(0,v)$. Indeed, in the other two cases ($(1,1)$ or $(1,v)$), observations from online political participation identify $\tilde{\mu}$, implying that $\mathbb{E}_\chi[\mu] = \tilde{\mu}$. This would contradict the second row of System 7. Therefore, on top of system 7, any candidate learning trap $(a, \chi, (\tilde{\lambda}, \tilde{\mu}))$ must satisfy a second system of inequalities that guarantees that the online equilibrium is $(0,1)$ or $(0,v)$. In addition, it is easy to see that this second system is not implied by the first: take for instance parameters such that System 7 is satisfied, but

$$\begin{cases} v\theta_R + (\gamma - \alpha)\tilde{\lambda}\tilde{\mu} < \kappa_v[\overline{c} - \underline{c}], \\ \kappa_1\underline{c} < \theta_M + \alpha\mathbb{E}_\chi[\mu]. \end{cases}$$

Such combination of parameters clearly exists, and is such that the online equilibrium played is $(1,1)$. This proves the first part of Proposition 2.

To prove the second part, note that, if κ_1 and κ_v are small enough, then the only online equilibrium is $(1,v)$. This implies that a self-confirming equilibrium must satisfy $\chi = \delta_{\tilde{\lambda},\tilde{\mu}}$, and hence no learning trap is possible. ∎

A.4 Learning Traps Caused by Social Media

We here show that, for fixed prior, social media might hinder learning about μ, yielding a learning trap where moderates underestimate their share. We formulate this claim in a special case in Proposition 5:

Proposition 5 *Fix $n = +\infty$. There exist parameter values and prior distributions χ_0 and χ_0' such that:*

(i) Under belief χ_0, the game without social media converges on the full-information equilibrium $(1,1)$, whereas the game with social media converges on $(0,v)$.

(ii) Under belief χ_0', the game without social media converges on the full-information equilibrium $(1,1)$, whereas the game with social media converges on $(0,1)$.

We prove item (i) by exhibiting parameters such that the full-information equilibrium $(1,1)$ is played forever in the game without social media, while $(0,v)$ is played forever in the game with social media.

A set of sufficient conditions is given by equations 8 to 15:

$$\begin{cases} \theta_M + \alpha \mathbb{E}_{\chi_0}[\mu] > \underline{c}, & (8) \\ \theta_M + \alpha \tilde{\mu} > \underline{c}, & (9) \\ \theta_M + \alpha \mathbb{E}_{\chi_0}[\mu | \tilde{\lambda}\tilde{\mu}] > \underline{c}, & (10) \\ v\theta_R + \gamma \mathbb{E}_{\chi_0}[\lambda\mu] - \alpha \mathbb{E}_{\chi_0}[\mu] < \overline{c} - \underline{c}, & (11) \\ v\theta_R < \kappa_v[\overline{c} - \underline{c}], & (12) \\ v\theta_R + \gamma\tilde{\mu}\mathbb{E}_{\chi_0}[\lambda|\tilde{\mu}] - \alpha\tilde{\mu} < \overline{c} - \underline{c}, & (13) \\ v\theta_R + \gamma \mathbb{E}_{\chi_0}[\lambda\mu] - \alpha \mathbb{E}_{\chi_0}[\mu] > \kappa_v[\overline{c} - \underline{c}], & (14) \\ v\theta_R + \gamma\tilde{\lambda}\tilde{\mu} - \alpha \mathbb{E}_{\chi_0}[\mu \mid \tilde{\lambda}\tilde{\mu}] > \overline{c} - \underline{c}, & (15) \end{cases}$$

It is easy to check that this system admits some solutions. Note how Condition (15) implies that the realized value of $\tilde{\lambda}$ is large relative to the ex-ante expectation.

Let us analyze the game without social media. Conditions (8), (11) and (12) (which implies $v\theta_R < \overline{c} - \underline{c}$), imply that $(1,1)$ is played in the first period, revealing $\tilde{\mu}$. Conditions (9) and (13) then imply that $(1,1)$ is played in every period thereafter.

Let us now analyze the game with social media, and show that $(0,v)$ can be played in every period after being selected against $(1,1)$ in the "indeterminate region" of Figure 1, which is non-empty at every period by virtue of combining condition (12) with (11) and (15). Conditions (8) and (14) imply that $(0,v)$ is played at the first period. This reveals the value of $\tilde{\lambda}\tilde{\mu}$. Then conditions (10) and (15) imply that $(0,v)$ is played at any subsequent period, online or offline. This concludes the proof of item (i).

We prove item (ii) similarly by exhibiting parameters such that the full-information equilibrium $(1,1)$ is played forever in the game without social media, while $(0,v)$ is played in the first period of the game with social media, followed by $(0,1)$ forever. In the system of conditions (8)-(15) (substituting χ_0' for χ_0 everywhere), we replace Condi-

tions (10) and (15) by

$$\begin{cases} \theta_M + \alpha \mathbb{E}_{\chi_0'}[\mu | \tilde{\lambda}\tilde{\mu}] < \underline{c}, & (16) \\ v\theta_R + (\gamma - \alpha)\tilde{\lambda}\tilde{\mu} < \kappa_v[\overline{c} - \underline{c}] & (17) \end{cases}$$

Again, one can check that this system admits some solutions. Note how Condition (17) now implies that the realized value of $\tilde{\lambda}\tilde{\mu}$ is small relative to the ex-ante expectation. In the game with social media where $\tilde{\lambda}\tilde{\mu}$ is revealed at the first period, this prompts a shift to the equilibrium $(0, 1)$ both online and offline thereafter. ∎

A.5 Proof of Proposition 3

Observe first that each sequence reveals $\tilde{\mu}$ in period 1a and $\tilde{\lambda}$ in period 2a. So the populations' beliefs at the beginning of the period are χ_0 in 1a, $\mathbb{E}_{\chi_0}[. \mid \tilde{\mu}]$ in 1b and 2a, and $\delta_{\tilde{\lambda},\tilde{\mu}}$ from period 2b on.

The fact that $(1, 1)$ is played in period $t = 1b$ shows that $\theta_M + \alpha\tilde{\mu} \geq \underline{c}$, which precludes the equilibrium $(0, 1)$ from period 2a on. Therefore there are only three possibilities from period 2b, as stated in the proposition.

We now exhibit parameter values for which these three sequences occur. The parameter values are the same for all sequences except for the realizations $\tilde{\lambda}_1, \tilde{\lambda}_2$ and $\tilde{\lambda}_3$ of λ. Take $\tilde{\lambda}_3 \approx 1$, $\tilde{\lambda}_1 \approx 0$, $\mathbb{E}_{\chi_0}[\mu] \approx 0$. Then one can find parameter values that satisfy all the following conditions:

$$\begin{cases} \theta_R < \underline{c}, & (18) \\ \theta_M + \alpha\tilde{\mu} - (\alpha + \beta)\tilde{\lambda}_2\tilde{\mu} > \underline{c}, & (19) \\ \theta_M - \beta\tilde{\mu} < \kappa_1\underline{c}, & (20) \\ \theta_M > \kappa_1\underline{c}, & (21) \\ \theta_M + \alpha\tilde{\mu} - (\alpha + \beta)\mathbb{E}[\lambda \mid \tilde{\mu}] > \kappa_1\underline{c}, & (22) \\ v\theta_R < \kappa_v(\overline{c} - \underline{c}), & (23) \\ v\theta_R + (\gamma - \alpha)\mathbb{E}[\lambda \mid \tilde{\mu}]\tilde{\mu} < \overline{c} - \underline{c}, & (24) \\ v\theta_R + (\gamma - \alpha)\mathbb{E}[\lambda \mid \tilde{\mu}]\tilde{\mu} > \kappa_v(\overline{c} - \underline{c}), & (25) \\ v\theta_R + (\gamma - \alpha)\tilde{\lambda}_2\tilde{\mu} > \overline{c} - \underline{c}. & (26) \end{cases}$$

We now explain how these conditions imply the sequences in Figure 4. First, observe that $\theta_R < \underline{c}$ (condition 18) and $v\theta_R < \kappa_v(\overline{c} - \underline{c}) < \overline{c}$ (condition 23) imply that radicals (and a fortiori moderates) play $a = 0$ forever in the absence of social media (given $\mathbb{E}_{\chi_0}[\mu] \approx 0$).

Second, with social media, the inequality $\theta_M > \kappa_1 \underline{c}$ (condition 21) together with $v\theta_R < \kappa_v(\overline{c} - \underline{c})$ (condition 23) implies that $(1,1)$ is played in period 1a. Inequality 19 then implies $\theta_M + \alpha\tilde{\mu} > \underline{c}$; together with inequality 24, this implies that $(1,1)$ is also played in period 1b. For period 2a, inequalities 25 and 22 imply that $(1,v)$ is played. Last, from period 2b on, in the first sequence (with $\tilde{\lambda}_1 \approx 0$), condition 23 implies that $(1,1)$ is played, in the second sequence conditions 19 and 26 imply that $(1,v)$ is played, and in the third sequence (with $\tilde{\lambda}_3 \approx 1$) conditions 20 and 26 (which implies $v\theta_R + (\gamma - \alpha)\tilde{\mu} > \overline{c} - \underline{c}$) guarantee that $(0,v)$ is played.

A.6 Proof of Proposition 4

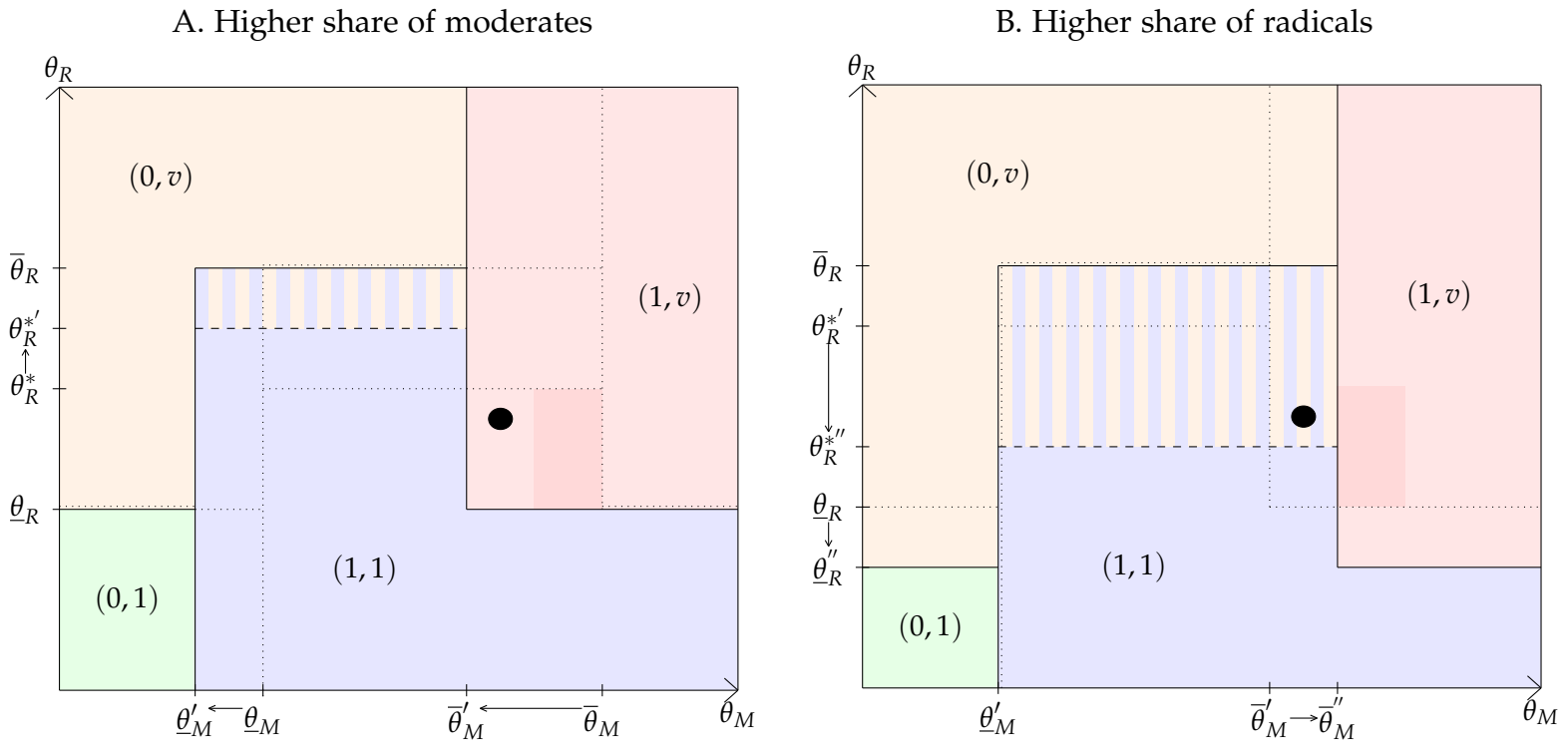
To prove item (i), note that, for each sequence, $(0,0)$ being the offline equilibrium in the absence of social media in period 1b implies $\theta_M + \alpha\mathbb{E}_{\chi_0}\mu < \underline{c}$, whereas $(1,1)$ being the offline equilibrium with social media indicates that $\theta_M + \alpha\tilde{\mu} \geq \underline{c}$. Hence, $\tilde{\mu} > \mathbb{E}_{\chi_0}\mu$. Note also that $v\theta_R \leq \kappa_v(\overline{c} - \underline{c}) < \overline{c} - \underline{c}$, otherwise the initial equilibrium would feature $a = v$ by the radicals. Together with $\theta_M + \alpha\tilde{\mu} \geq \underline{c}$, this implies that, in the crowd-in-then-crowd-out-sequence, the periods at which $(0,v)$ is played correspond to the striped region in Figure 1.

To prove item (ii), let us focus on the third sequence and compare the equilibra in periods 1a and 3a. The fact that $(0,v)$ is played in period 3a implies that $v\theta_R + \gamma\tilde{\lambda}_3\tilde{\mu} - \alpha\tilde{\mu} > \kappa_v(\overline{c} - \underline{c})$, whereas the fact that $(1,1)$ is played in period 1a implies that $v\theta_R + \gamma\mathbb{E}_{\chi_0}[\lambda\mu] - \alpha\mathbb{E}_{\chi_0}[\mu] \leq \kappa_v(\overline{c} - \underline{c})$. Combining these expressions yield $\gamma\tilde{\lambda}_3\tilde{\mu} - \alpha\tilde{\mu} > \gamma\mathbb{E}_{\chi_0}[\lambda\mu] - \alpha\mathbb{E}_{\chi_0}[\mu]$ and, since $\tilde{\mu} > \mathbb{E}_{\chi_0}[\mu]$, $\tilde{\lambda}_3\tilde{\mu} > \mathbb{E}_{\chi_0}[\lambda\mu]$.

Let us now compare the equilibrium in period 3a with that in period 2a. That $(0,v)$ is played in period 3a while $(1,v)$ is played in period 2a implies $\theta_M + \alpha(1 - \tilde{\lambda}_3)\tilde{\mu} - \beta\tilde{\lambda}_3\tilde{\mu} \leq \kappa_1\underline{c}$ and $\theta_M + \alpha(1 - \mathbb{E}_{\chi_0}[\lambda \mid \tilde{\mu}])\tilde{\mu} - \beta\mathbb{E}_{\chi_0}[\lambda \mid \tilde{\mu}]\tilde{\mu} > \kappa_1\underline{c}$, from which we infer $\tilde{\lambda}_3 > \mathbb{E}_{\chi_0}[\lambda \mid \tilde{\mu}]$. This establishes item (ii).

To prove item (iii), note that $(1,1)$ and $(0,v)$ being the equilibria for sequences 1 and 3 respectively in period 2b implies that $v\theta_R + \gamma\tilde{\lambda}_3\tilde{\mu} - \alpha\tilde{\mu} > \overline{c} - \underline{c} \geq v\theta_R + \gamma\tilde{\lambda}_1\tilde{\mu} - \alpha\tilde{\mu}$, from which we infer $\tilde{\lambda}_3 > \tilde{\lambda}_1$. Similarly, $(1,v)$ and $(0,v)$ being the equilibria for sequences 2 and 3 respectively in period 2b implies that $\theta_M + \alpha(1 - \tilde{\lambda}_2)\tilde{\mu} - \beta\tilde{\lambda}_2\tilde{\mu} \geq \underline{c} > \theta_M + \alpha(1 - \tilde{\lambda}_3)\tilde{\mu} - \beta\tilde{\lambda}_3\tilde{\mu}$, which yields $\tilde{\lambda}_3 > \tilde{\lambda}_2$.

Figure A.1: Comparative statics on population shares and protest dynamics



Notes: The striped area corresponds to equilibrium $(1,1)$ if $\gamma\mathbb{E}[\lambda\mu] < \alpha\mathbb{E}[\mu]$. In that case, θ_R^* is not defined. Conversely, if $\gamma\mathbb{E}[\lambda\mu] > \alpha\mathbb{E}[\mu]$, the equilibrium is indeterminate, with moderates preferring $(1,1)$ and radicals preferring $(0,v)$. The dotted (respectively, solid) lines delimit the equilibria before (respectively, after) the revelation of a higher share of moderates (Panel A), and before (respectively, after) the subsequent revelation of a higher share of radicals while keeping the share of passives fixed (Panel B). We also represent two possible dynamics following a $(1,1)$ equilibrium. In the dark red area, after the revelation of a higher share of moderates, the equilibrium switches from $(1,1)$ to $(1,v)$, but remains stable after the subsequent revelation of a higher share of radicals. Conversely, the black dot locates a region compatible with the crowd-in-then-crowd-out sequence, where $(1,1)$, $(1,v)$ and $(0,v)$ are played sequentially provided radicals are able to impose their preferred equilibrium $(0,v)$ in the indeterminate region.

B Elements of Context

In 2015, then-President François Hollande decided to gradually introduce a carbon tax on top of the existing gas tax in order to converge the after-tax price of diesel and gasoline. The carbon tax was confirmed in 2017 by the newly elected President Emmanuel Macron, despite the fact that oil prices had been rising since 2016 and that car-related expenses had been increasing for several years. A few months later, in January 2018, Prime Minister Philippe decided to lower the speed limit on secondary roads from 90 km/h to 80 km/h, citing concerns about road safety. This latter decision, which was not part of Emmanuel Macron's campaign manifesto, led to the organization of numerous slowdowns across the country. The new 80 km/h regulation came into force on July 1, 2018.

At the end of the summer recess, the annual increase in the carbon tax was confirmed in the 2019 budget, despite growing discontent, especially among motorists. A few months earlier, in May 2018, a motorist had started a petition against the gas tax on the Change.org platform. Although the petition had only received a few hundred signatures in its first few months, it was mentioned in a local newspaper on October 12, 2018. This newspaper had a local readership in *Seine-et-Marne* (a county on the outskirts of the Paris region), where the article triggered a first wave of signatures. The wife of a truck driver who planned to block the Paris ring road in November for 17th read the article and linked 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 roadblock project, and signatures skyrocketed nationwide. On October 24, an online video suggested the yellow safety vest, which all car owners are required by law to have in their trunks, as a rallying sign for angry drivers. Roadblock organizers relied heavily on Facebook to spread the word, and several dedicated 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

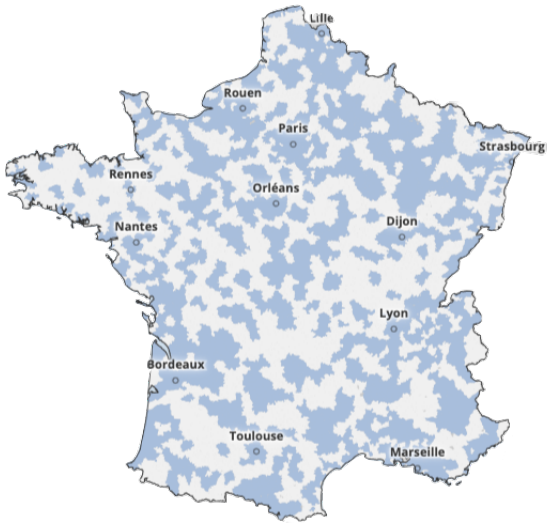
A website (www.blocage17novembre.fr) was created to coordinate the mobilization. It provided a map of the organized blockades, updated in real-time. As of November 16, the map documented 788 geolocated blockades. We use this map to document the offline mobilization of the Yellow Vests, summarized in Figure C.1.

Starting from January 19th, 2019 (the seventh week of the Yellow Vest movement), a group of Yellow Vests, called *Le Nombre Jaune* ("the Yellow Number") started to collect statistics about the number of participants to Yellow Vest demonstrations across the country. Each week, they published a dataset containing a list of Yellow Vest demonstrations that took place on that week's Saturday, along with the estimated number of demonstrators that participated in each protest. To build these datasets, they relied on articles from local newspapers, videos published online, as well as reports from protesters. In Figure C.2, we use these statistics to show measures of the number and size of these protests until May 2019.

Figure C.1: Blocking Half of France at First Try

A. Geolocation

B. Affected Living Zones



Notes: Panel A displays the geolocation of the 11/17 roadblocks. Panel B displays the living zones with at least one roadblock on 11/17. These living zones gather 49 million people, 77% of the French mainland population.

C.2 Change.org Petition

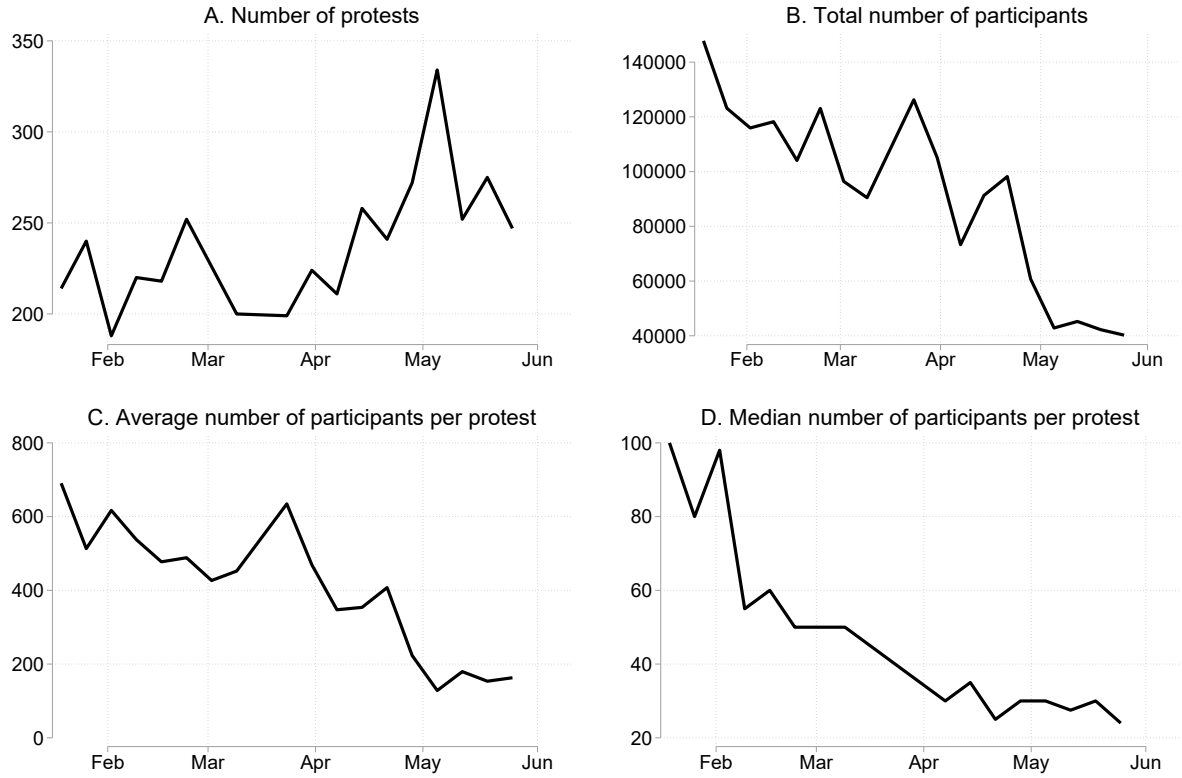
Change.org gave us access to an anonymized list of the signatories of the petition "Pour une baisse du prix des carburants à la pompe". Each observation is associated with the date of signature and the ZIP code of the signatory. We restrict the data to signatures in mainland France and with a valid ZIP 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. In Figure C.3, we map the distribution of signature rates 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 be-

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

Figure C.2: Measures of offline Yellow Vest activity from *Le Nombre Jaune*



Notes: This figure describes the frequency and magnitude of Yellow Vest protests in the first four months of 2019, as reported by *Le Nombre Jaune*. We do not report numbers for March 16th, 2019, when the Yellow Vest protests were organized jointly with a demonstration for climate awareness ("marche pour le climat").

tween 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

[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. Figure C.4 displays the spatial distribution of these groups before (Panel A) and after (Panel B) 11/17.

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.

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.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.

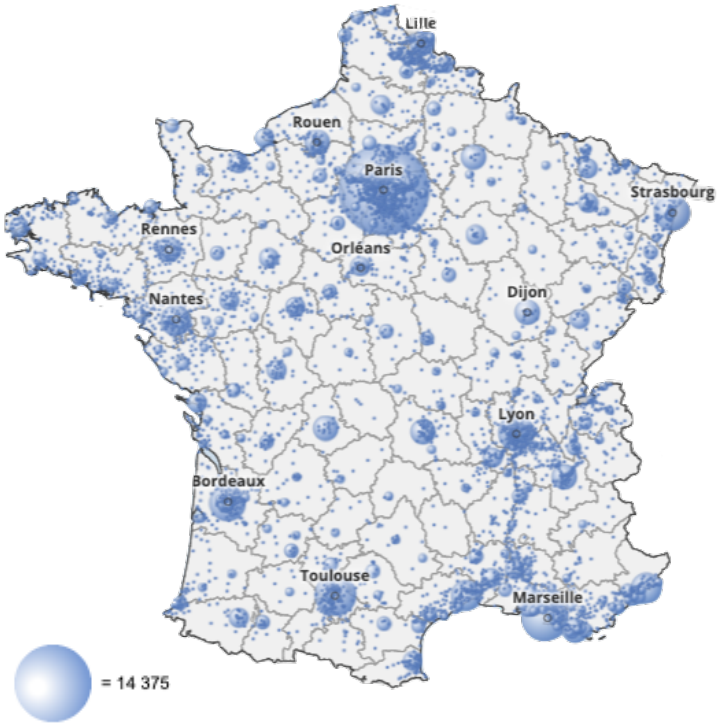
Table C.2: Comparison Between the Two Data Collections on Facebook Pages

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

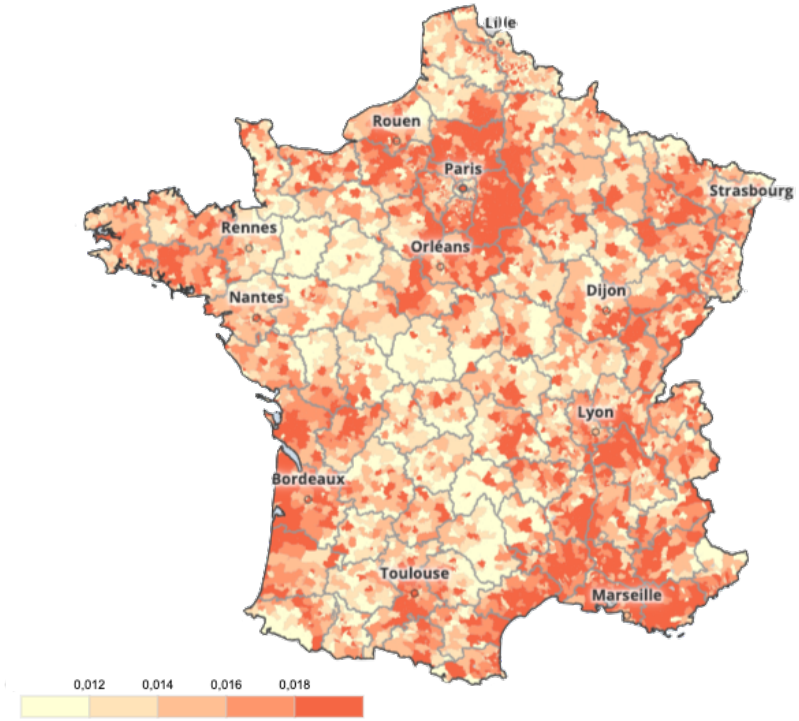
Notes: This table presents simple count metrics to compare the datasets resulting from our two data collections on Facebook pages.

Figure C.3: Signature Rate of the Change.org Petition per Municipality

A. Absolute value



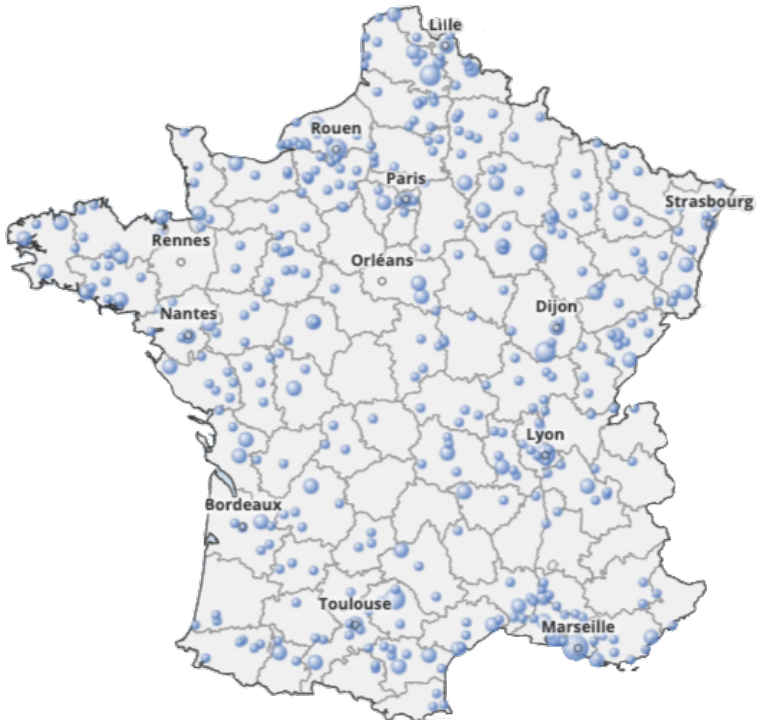
B. Per inhabitant



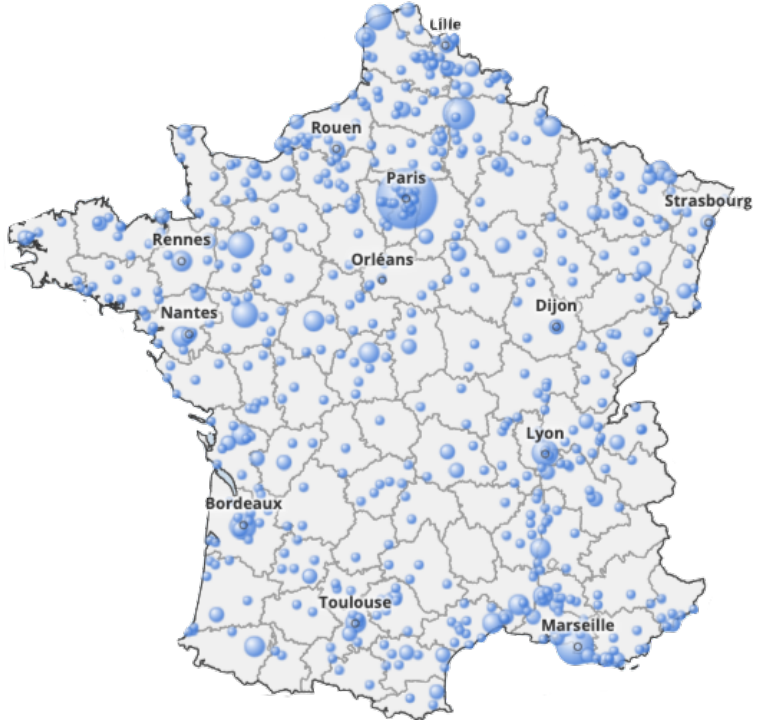
Notes: Figure A displays the number of signature per municipality. Figure B displays the signature rate (signature per inhabitant) by municipality.

Figure C.4: Number of Local Groups per Municipality.

A. Before 17/11



B. After 17/11



Notes: The two figures display the number of Yellow Vests local groups per municipality. Figure A corresponds to group creation before 11/17, while Figure B corresponds to group creation after 11/17.

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 Administrative data at the municipal level

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

Control variables.

- **Geography** includes the population of the municipality (we also include its square and two splines), its density, its altitude, the distance to the closest city with over 20,000 inhabitants and 100,000 inhabitants, whether the municipality was classified as urban in 2015, and whether it switched from rural to urban between 1999 and 2015. *Source: Census (RP, complementary exploitation), 2016, INSEE.*

- **Transport** 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.*

- **Economy** includes the local unemployment rate, the fraction of employees with a non-permanent contract, mean income, and population immigrant share. *Source: Census 2016, INSEE. DADS, 2015, INSEE.*

- **Occupation** includes the share of the different *catégories socio-professionelles* defined by INSEE: executive, independent, middle-management, employee, manual worker and agriculture. *Source: Census 2016, INSEE.*
- **Age** includes 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.*
- **Education** includes the shares of the population without a high-school diploma, and with a university degree. *Source: Census 2016, INSEE.*
- **Vote** includes the vote share for the five major candidates in the 2017 presidential election (Macron, Le Pen, Fillon, Mélenchon, Hamon), as well as the share of abstention.*Source: Ministry of the Interior.*
- **LZ** is a set of dummies for Living Zones. *Source: INSEE.*

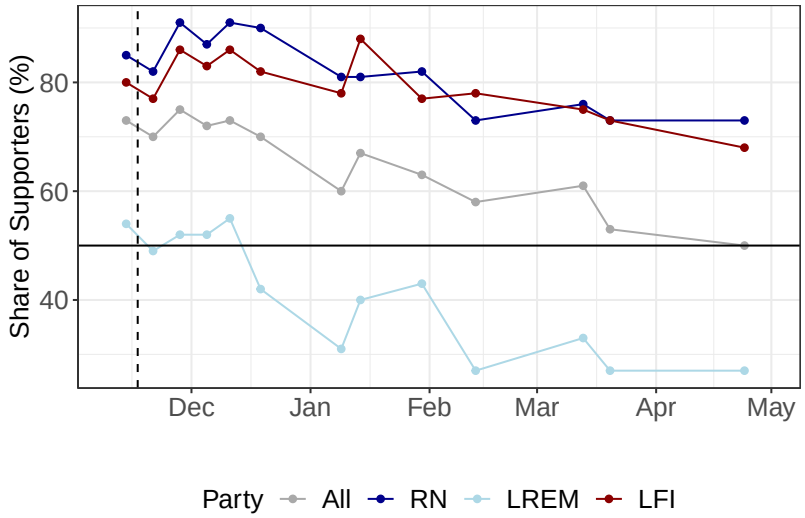
Instruments.

- **Roundabouts** is the number of roundabouts per square kilometer in the municipality and in the other municipalities of the Living Zone. *Source: OpenStreetMap.*
- **4G Coverage** measures exposure to 4G as the log number of days since the installation of a 4G antenna prior to 11/17. *Source: Agence Nationale des Fréquences.*

C.6 Polls

The polling institute ELABE conducted several surveys between November 2018 and April 2019 for the news Channel BFM TV. Figure C.5 reports their results on the evolution of popular support for the Yellow Vests movement.

Figure C.5: Evolution of the Popular Support for the Yellow Vests

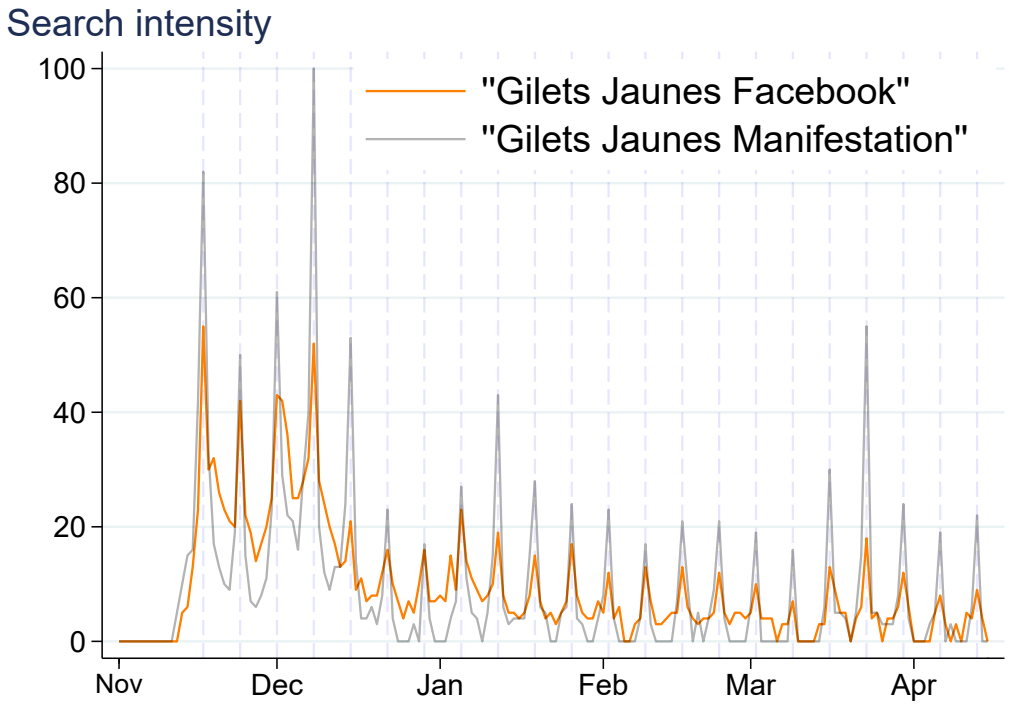


Notes: This figure plots the share of the population who declared they were supportive or sympathetic to the Yellow Vests movement over time. The vertical dashed line corresponds to 11/17. ELABE, the survey institute from which we collected data, conducted polls on 11/14/2018, 11/21/2018, 11/28/2018, 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.7 Google Trends

Figure C.6 shows daily statistics from Google Trends in France for two phrases: *Gilets Jaunes Facebook* and *Gilets Jaunes Manifestation*. Street protests 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.6: Evolution of Google searches

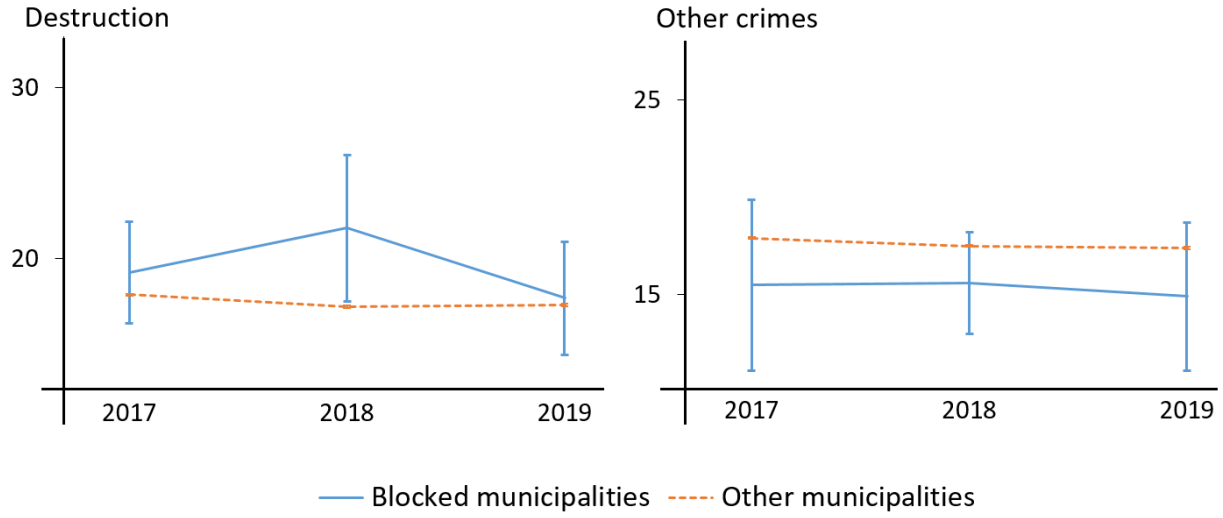


Notes: Daily index of Google Search intensity in France for the keywords *Gilets jaunes Facebook* and *Gilets jaunes Manifestation* between November 1st, 2018 and April 15th, 2019. The dashed lines correspond to the weekly protests, starting in 11/17. *Source*: Google Trends.

C.8 Street violence

To construct a measure of street violence, we use official data from the Ministry of the Interior, which provides counts of offenses recorded by the police. This data is either available as a monthly panel at the regional level or as a yearly panel at the municipal level. We isolate one class of offenses: "destruction of public and private property", which we use as a proxy for rioting. We also construct a "placebo" measure that includes offenses related to other criminal activities (vehicle theft and drug trafficking). We measure net crime in a given year as the prediction error from a regression of the variable of interest in year t on its value in year $(t-1)$, the value of the other offense in year $(t-1)$, our set of municipal covariates, and the Living Zone fixed effects. As shown in Figure C.7, the average level of destruction in municipalities that were blocked in 11/17 was 30% higher than in other municipalities in 2018. Conversely, the difference is smaller and not statistically significant for 2017 and 2019, or for the other crime category in 2018.

Figure C.7: Crime in blocked municipalities and other municipalities



Notes: Net measure of crime (with 90% confidence intervals) in municipalities that were blocked on 11/17 and in other municipalities. In addition to the value of both offenses in the previous year, the value is net of local characteristics and Living Zone fixed effects. The list of controls is detailed in Appendix C.5.

D Supplement for the municipal analysis

D.1 IV results on the impact of early online mobilization on protests

To retrieve information on the distribution of 4G Antennas, we use up-to-date (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. In 2024, more than 15,000 municipalities had their own 4G antenna, against less than 9,000 before 11/17.

We define our instrument as the (log) number of days since the installation of the first 4G antenna in the municipality prior to 11/17. However, some antennas cover multiple municipalities, making coverage difficult to predict accurately because it does not only depend on distance to the closest antenna, but also on other local characteristics as well as the antenna's technology or the fact that it may or may not be shared between operators. Therefore, we restrict the analysis to the municipalities that eventually received an antenna, for which we can more safely assume that signal quality was weaker before the installation of an antenna. In addition, to improve the comparability between early-treated and later-treated municipalities, we restrict the sample to all municipalities that received at least one antenna after 11/17.

The identifying assumption behind the use of the 4G variable as an instrument is that conditional on observable characteristics, the timing of 4G roll-out only predicts the organization of roadblocks through its impact on early Yellow Vest early online activity. Admittedly, the roll-out of 4G was not random and started with more dense areas (see Appendix Figure D.1-a): geography, as described in Section C.5, explains over 30% of the spatial variation in the installation date of the first antenna. However, the roll-out was partly driven by operational constraints, in particular the organization of frequency auctions by the government (in 2011 and 2015). After residualizing the installation date with all our control variables except geography, geographical characteristics related to altitude, population, density or urbanization only explain 3% of the remaining variation. Visually, once we control for our full set of control variables, 4G access on the onset of the Yellow Vest movement appears randomly distributed—see Appendix Figure D.1(b).

The second difficulty stems from the fact that many unobservable social media were likely used to organize the 11/17 roadblocks. We can partly circumvent this problem by using two alternative means of online mobilization, measured by the petition signature rate and the number of early Facebook groups. We combine both variables by isolating municipalities that belong to the fourth quartile of either variable, then defining two

Table D.1: Effects of Early Online Activity on the 11/17 roadblocks

	Probability of roadblock					
	Instrumental variable				**OLS**	
	High mobilization		**Top mobilization**			
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A:	**2SLS**					
Online Mobilization	0.357***	0.343***	0.733***	0.670***		
	(0.081)	(0.024)	(0.207)	(0.055)		
Panel B:	**First stage**				**Reduced-form**	
4G coverage	0.008***	0.038***	0.004***	0.020***	0.003***	0.013***
	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)
Controls	✓		✓		✓	
Living Zone FE	✓		✓		✓	
Observations	13,603	13,603	13,603	13,603	13,603	13,603
Kleibergen-Paap F-stat	39.2	292.5	17.1	166.3		
R-Squared					0.316	0.041

Notes: Except in columns 5 and 6, this table shows 2SLS estimates of the impact of early online activity on the probability of a roadblock on 11/17, instrumented by the log number of days since the first 4G antenna was installed in the municipality before 11/17 if no additional antenna was installed after 11/17 and zero otherwise. In columns 1 and 2, online mobilization is defined as a dummy variable equal to 1 if the municipality belongs to the top quartile in terms of the early petition signature rate or in terms of the early number of Facebook groups. In columns 3 and 4, the online mobilization is defined as a dummy variable equal to 1 if the municipality belongs to the top quartile in terms of the early petition signature rate and in terms of the early number of Facebook groups. Coverage 4G shows the first-stage estimates in columns 1 to 4 and the reduced-form estimates in columns 5 and 6. The sample is restricted to all municipalities that received a 4G antenna between 11/17 and May 2024. Estimates of the first stage are shown for the instrument. We cluster standard errors at the Living Zone level. *: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$.

dummy variables coding for either the union or the intersection of those two characteristics. The former corresponds to 60% of municipalities in the regression sample, and allows for substitutability between the two platforms. The latter corresponds to 14% of municipalities in the regression sample and rests upon the assumption that both platforms are complementary. We label them "High mobilization" and "Top mobilization", respectively.

Table D.1 presents the regression results. The instrument is positively correlated with both the regressor (columns 1 to 4) and with the outcome variable (columns 5 and

Table D.2: Effects of Early Online Activity on the 11/17 roadblocks: Robustness

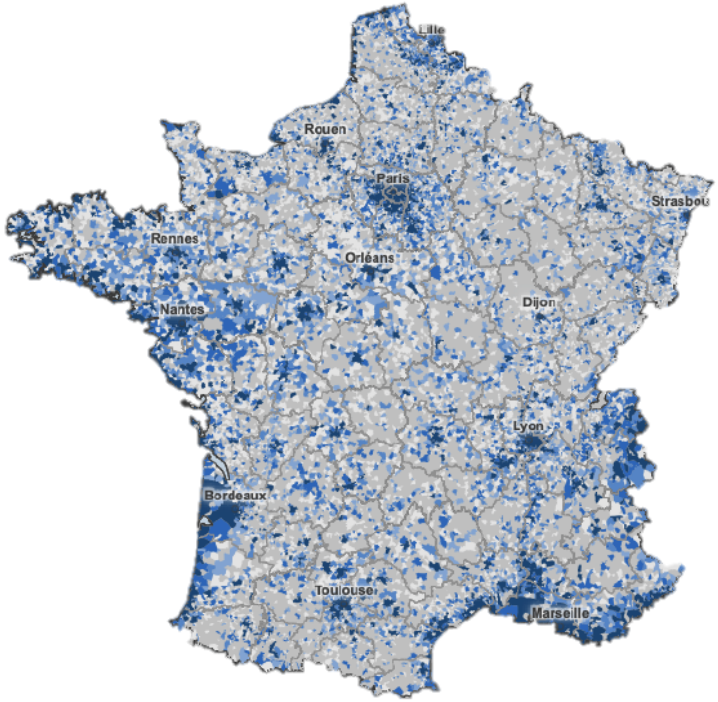
	Probability of roadblock					
	High mobilization			**Top mobilization**		
	(1)	(2)	(3)	(4)	(5)	(6)
	2SLS					
Online Mobilization	0.187***	0.293***	0.303***	0.408***	0.643***	0.640***
	(0.026)	(0.078)	(0.082)	(0.059)	(0.210)	(0.207)
Controls	✓	✓	✓	✓	✓	✓
Living Zone FE	✓	✓	✓	✓	✓	✓
Full sample	✓			✓		
Binary instrument		✓			✓	
Excluding Paris region			✓			✓
Observations	34,434	13,603	12,835	34,434	13,603	12,835
Kleibergen-Paap F-stat	247.7	31.8	33.3	112.2	13.2	15.7

Notes: This table shows the 2SLS estimates displayed in Table D.1 under alternative specifications. In Columns 1 and 4, the sample is extended to all municipalities and the instrument is given a value equal to zero to all never-treated municipalities. In Columns 2 and 5, the instrument is a dummy variable equal to 1 if a 4G antenna was installed before 11/17. In Columns 3 and 6, we exclude the Paris region. In columns 1 to 3, online mobilization is defined as a dummy variable equal to 1 if the municipality belongs to the top quartile in terms of the early petition signature rate or in terms of the early number of Facebook groups. In columns 4 to 6, online mobilization is defined as a dummy variable equal to 1 if the municipality belongs to the top quartile in terms of the early petition signature rate and in terms of the early number of Facebook groups. We cluster standard errors at the Living Zone level. *: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$.

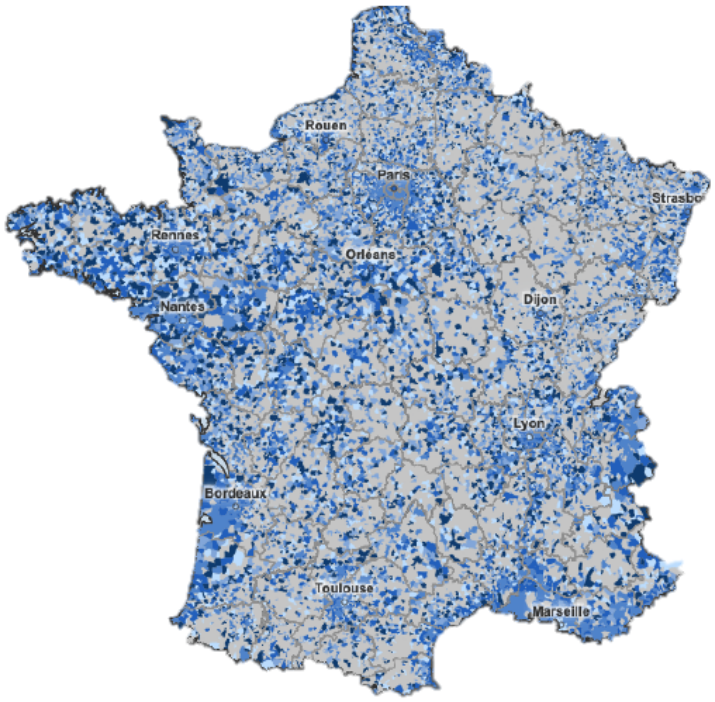
6). The Kleibergen-Paap F-statistic is equal to 39 in column (1), suggesting that our instrument is reasonably strong for our first proxy of online mobilization. It is slightly lower in column (3), although still above the threshold value of 16.4 for a maximal size of 10% provided by Stock and Yogo (2005). Second-stage estimates are quite similar in specifications without controls (columns 2 and 4), which is reassuring regarding the validity of our exclusion restriction. As is expected, the estimates are also higher in columns (3) and (4), where online mobilization features both high petition signature rate and high number of Facebook groups. As shown in Table D.2, the results are robust to using the full sample of municipalities and assign a value of zero of the instrument to all never-treated municipalities. They are also robust to using a simple dummy variable coding for the presence of a 4G antenna in the municipality prior to 11/17 or to dropping the Paris region, which stands out along many dimensions.

Figure D.1: Rollout of 4G on 11/17

A. 4G coverage on 11/17



B. Residuals



Notes: Panel A shows exposure to 4G coverage on 11/17, as defined in the text. Panel B shows the residual of 4G coverage after controlling for the set of controls described in Section C.5. Color intensity corresponds to quantile thresholds. In grey, municipalities outside the regression sample.

D.2 IV Results on the impact of protests on later online mobilization

To instrument the roadblocks, we leverage the presence of roundabouts in each municipality. The rationale for the relevance of this instrument is that calls for demonstrations urged protesters to block roundabouts. By design, they allow to block several roads at a time and possess a central median strip on which it is convenient to set camp.The identifying assumption is that conditional on observable characteristics, the distribution of roundabouts only predicts future online mobilization through its impact on roadblocks. The history of roundabouts makes it likely that the conditional distribution of local roundabout density reflects local idiosyncrasies. Roundabouts are partly a French architectural fad, arguably invented in 1906 by the French urban planner Eugène Hénard. France has over sixty-thousand roundabouts (roughly four times more than the United Kingdom). One-third of French municipalities have at least one. While plausible road safety reasons support their use, they can almost always be replaced with traffic lights. In support of our exclusion restriction, a map of the prediction error of roundabout density after an OLS regression, including our controls, shows a seemingly random distribution (see Figure D.2).

Assuming the exogeneity of this first instrument, we can leverage a second instrument, which will allow us to test overidentifying restrictions. Indeed, since organizing a roadblock requires significant manpower, protesters had to coordinate to choose roadblock locations. This spatial coordination problem suggests another instrument, which is the mirror image of the first: the density of roundabouts in the other municipalities of the Living Zone. Because of competition between easy-to-block locations, we expect municipalities surrounded by more roundabouts to be less likely blocked.

Table D.3 presents the regression results. The Kleibergen-Paap F-statistic equals 25, suggesting that our instruments are reasonably strong. In addition, the high p-values associated with the Hansen J-statistics indicate that we fail to reject the hypothesis that the overidentifying restrictions are valid. The effect of each instrument goes in the expected direction, and may be directly observed in the reduced-form specification in columns (5) and (6). Column (1) shows that even though the bulk of petition signatures occurred before 11/17, having a roadblock increases the post-11/17 signature rate by 1.2 standard deviations. This result suggests that protests helped spread information about the Yellow Vests' demands at the end of 2018 when public support for the movement was still high. The previous signatory rate is also correlated with subsequent signatory dynamics.

We also find a strong positive impact of roadblocks on subsequent Facebook activity: a roadblock in a municipality increases the number of new local Facebook groups by 2.9 standard deviations (corresponding to 1.2 additional groups), which translates into

Table D.3: Effects of a Roadblock on Post-11/17 Online Mobilization

	Outcomes post-11/17					
		Facebook				
	Petition Signatures	Groups	Members	Posts	Petition Signatures	Facebook Groups
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A:	**2SLS**					
Roadblock	1.188***	2.957***	0.274***	0.187**		
	(0.254)	(0.695)	(0.097)	(0.079)		
Panel B:	**First-stage**				**Reduced-form**	
Roundabouts local	0.026***	0.026***	0.026***	0.026***	0.037***	0.093**
	(0.008)	(0.008)	(0.008)	(0.008)	(0.012)	(0.043)
Roundabouts LZ	-0.358***	-0.358***	-0.358***	-0.358***	-0.364***	-0.894**
	(0.073)	(0.073)	(0.073)	(0.073)	(0.117)	(0.413)
Controls	✓	✓	✓	✓	✓	✓
Living Zone FE	✓	✓	✓	✓	✓	✓
Pre-11/17 mobilization	✓	✓	✓	✓	✓	✓
Observations	34,434	34,434	34,434	34,434	34,434	34,434
Kleibergen-Paap F-stat	24.7	24.7	24.7	24.7		
p-value Hansen	0.570	0.533	0.286	0.255		
R-Squared					0.603	0.722

Notes: Except in columns (5) and (6), this table shows 2SLS estimates of the impact of a municipal roadblock on four measures of online mobilization after 11/17: the signature rate of the Change.org petition after 11/17 (column 1), the number of groups created post-11/17 (column 2), the number of members per inhabitant (column 3) and posts per inhabitant (column 4) in these newly created groups. Roundabouts local and roundabouts LZ show the first-stage estimates in columns (1) to (4) and the reduced-form estimates in columns (5) and (6). Roundabouts local is the number of roundabouts per square kilometer in the municipality. Roundabouts LZ is the number of roundabouts per square kilometer in all other municipalities of the Living Zone. Pre-11/17 mobilization is the signature rate pre-11/17 and the number of groups pre-11/17. Both outcome variables and instruments are standardized. We cluster standard errors at the Living Zone level. *: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$.

an increase in the number of new members per inhabitant by 0.21 standard deviations, and in the number of posts per inhabitant by 0.14 standard deviations. As shown in Table D.4, results are robust to several specification changes, such as using only one of the two instruments, not controlling for measures of early online mobilization, local characteristics or fixed effects, and dismissing the Paris region, which stands out along many dimensions.

Table D.4: Impact of Blockades on Post-17/11 Online Mobilization: Robustness

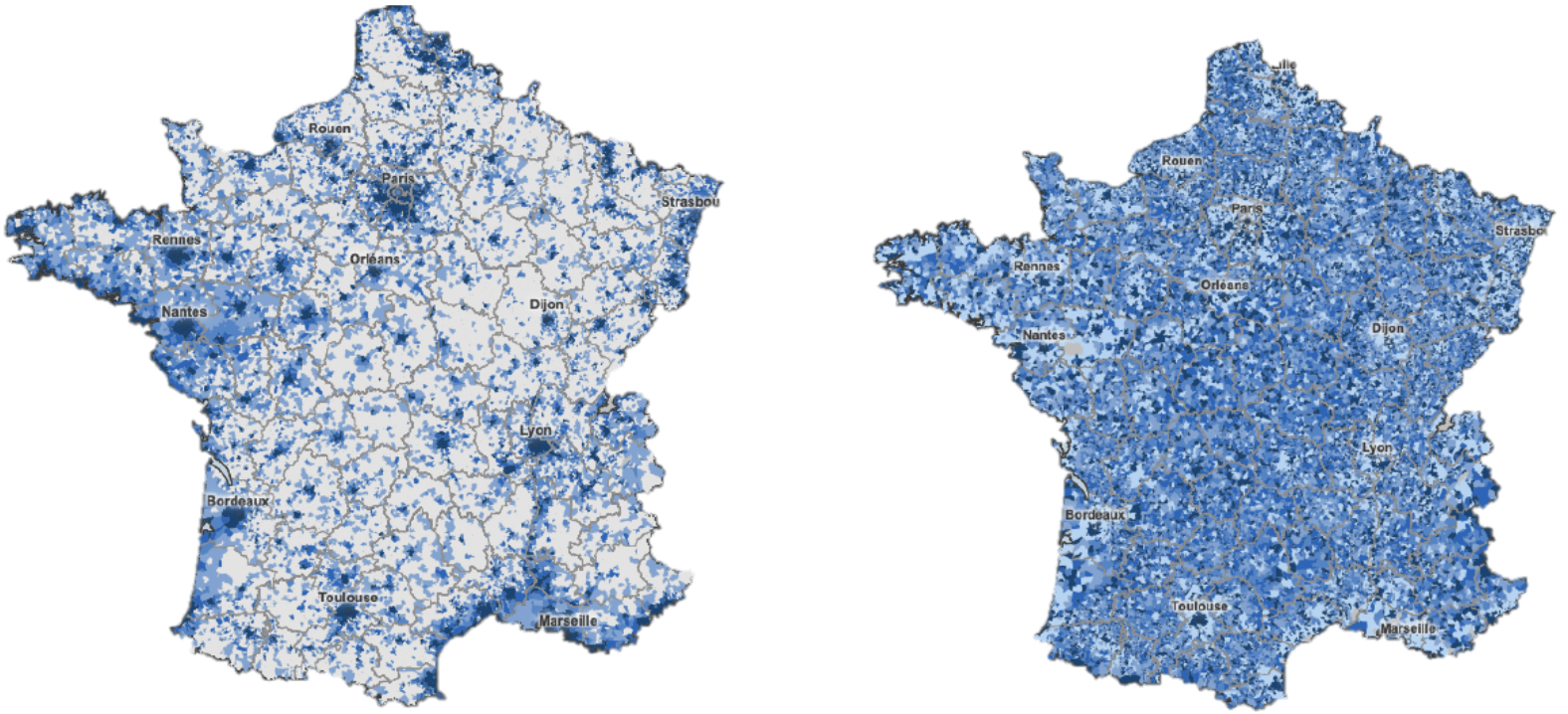
	Outcomes post-11/17			
	Petition Signatures	Facebook		
		Groups	Members	Posts
	(1)	(2)	(3)	(4)
Panel A: Without controls				
Blockade	2.740***	5.874***	0.155***	0.104***
	(0.283)	(0.490)	(0.051)	(0.039)
Kleibergen-Paap F-stat	30.6	30.6	30.6	30.6
p-value Hansen	0.001	0.022	0.036	0.029
Panel B: Only municipal instrument				
Blockade	1.327***	3.308***	0.321**	0.223*
	(0.350)	(0.874)	(0.157)	(0.117)
Kleibergen-Paap F-stat	14.6	14.6	14.6	14.6
Panel C: Only Living Zone instrument				
Blockade	1.097***	2.675***	0.169**	0.101*
	(0.305)	(0.843)	(0.076)	(0.059)
Kleibergen-Paap F-stat	37.0	37.0	37.0	37.0
Panel D: Commuting zone instead of Living Zone				
Blockade	0.635**	3.437***	0.294***	0.190**
	(0.254)	(0.898)	(0.097)	(0.075)
Kleibergen-Paap F-stat	12.1	12.1	12.1	12.1
p-value Hansen	0.099	0.872	0.128	0.137
Panel E: Excluding Paris region				
Blockade	0.871***	3.005***	0.410***	0.298**
	(0.309)	(0.904)	(0.152)	(0.132)
Kleibergen-Paap F-stat	18.9	18.9	18.9	18.9
p-value Hansen	0.718	0.625	0.064	0.071

Notes: This table shows the 2SLS estimates displayed in Table D.3 under alternative specifications. The municipal instrument is the number of roundabouts per square kilometer in the municipality. The living zone instrument is the number of roundabouts per square kilometer in all other municipalities of the Living Zone. In panel D, we replace Living Zones ($N = 1,631$) with Commuting Zones ($N = 297$) both for fixed effects and the definition of the leave-one-out instrument. Both outcome variables and instruments are standardized. We cluster standard errors at the Living Zone level, except in Panel D where we cluster them at the Commuting Zone level. *: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$.

Figure D.2: Roundabout density

A. Roundabouts by squared kilometer

B. Residuals

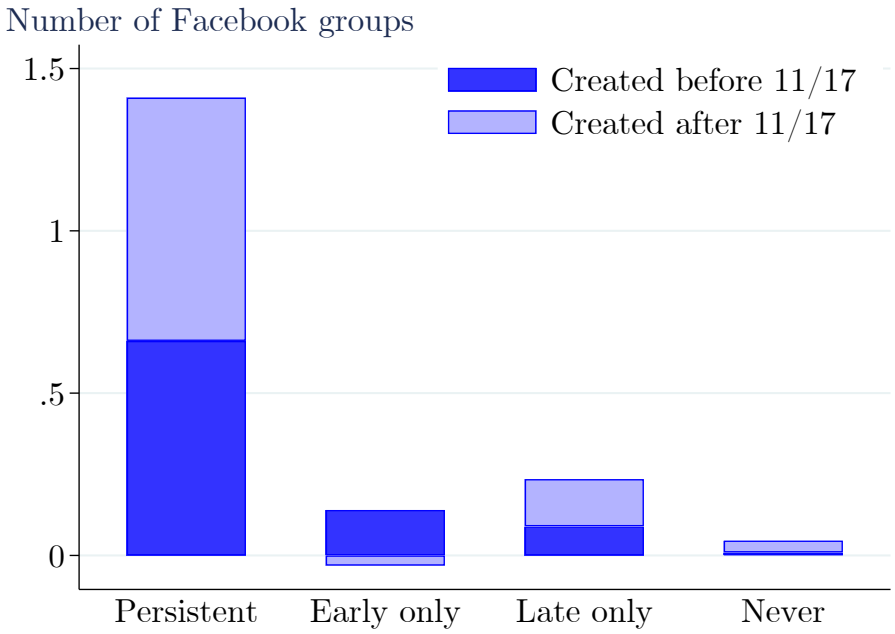


Notes: Panel A shows the density of roundabouts in mainland France, with darker colors corresponding to higher density. Panel B shows the residual density of roundabouts after controlling for the set of controls described in Section C.5. Color intensity corresponds to quantile thresholds.

D.3 Later mobilization

We use data from *Le Nombre Jaune* to further break down Figure 7 into four categories, based on the occurrence of a roadblock on 11/17 and at least one documented protest between January and May 2019. On average, municipalities that experienced persistent mobilization had four times more groups before 11/17 than municipalities where the 11/17 roadblocks did not translate into a protest in 2019 (the "Early only" category). However, this ratio rises to 1 for 11 for groups created after 11/17. As shown in Figure D.3, the difference is even more striking after controlling for local characteristics, Living Zone fixed effects and measures of pre-11/17 mobilization: the net number of groups created after 11/17 drops to zero in the group of early-only municipalities.

Figure D.3: Local Facebook groups and protest persistence



Notes: Average number of local Facebook groups in four categories of municipalities: those that experienced a roadblock on 11/17 and at least one protest between January and May 2019 according to the *Le Nombre Jaune* dataset (labelled as "persistent"); those that experienced a roadblock on 11/17 but no protest in 2019 (labelled as "early only"); those that were not blocked on 11/17 but experienced protests in 2019 (labelled as "late only"); and those that experienced neither. The value is net of local characteristics and Living Zone fixed effects. The list of controls is detailed in Appendix C.5. For groups created after 11/17, we also control for the number of groups created before 11/17 and for the petition signature rate before 11/17. Note that these later groups were all created before mid-december 2018 and therefore predate the 2019 protests.

E Supplement for "The two margins of online radicalization"

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).

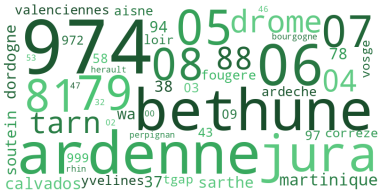
[7] We also considered alternative labeling options, such as term frequency-inverse cluster frequency, which yield similar results.

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%]

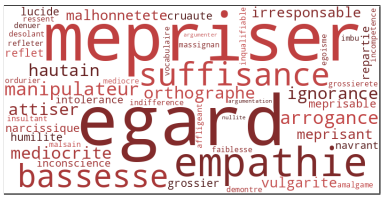
C. Support [3.9%]

D. Diffusion [4.8%]

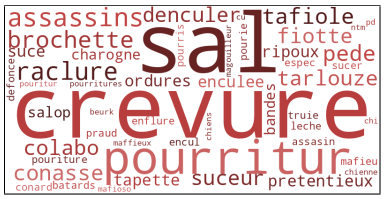
E. Economic concerns [5.9%]

F. Political institutions [7.6%]

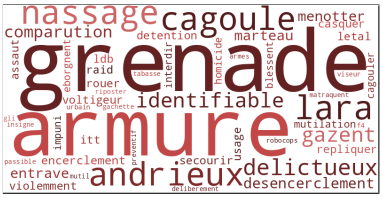
G. Food and objects [6%]



H. Critiques [6%]



I. Insults [4.5%]

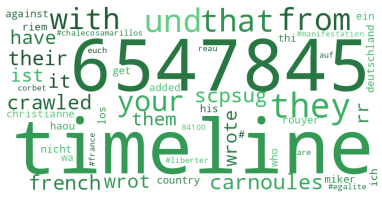


J. Violence [5.9%]

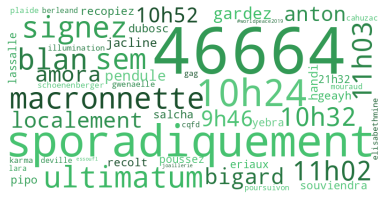
K. Conspiracy [5.9%]

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

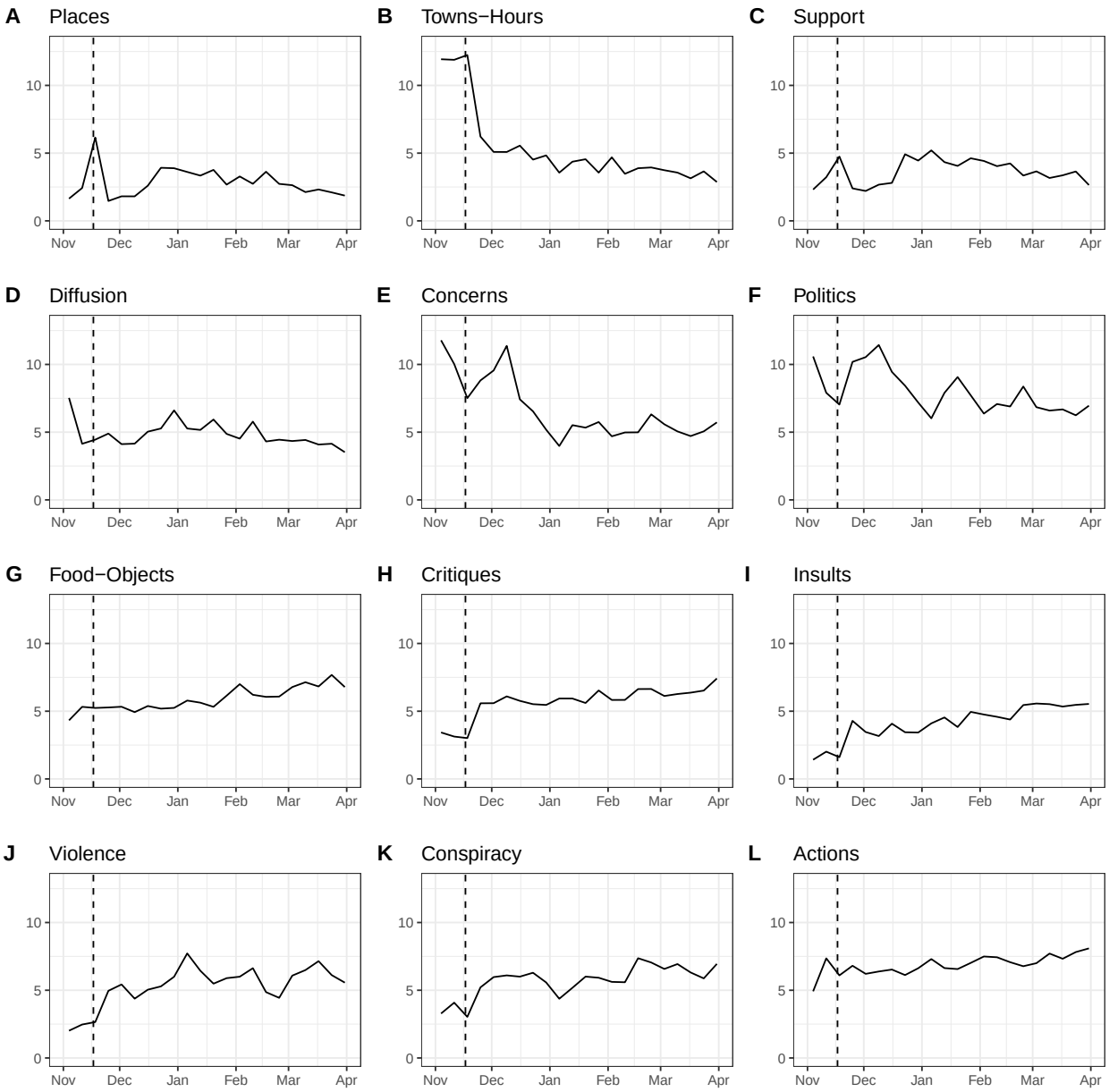
Il faut arreter de prendre des gants avec cette violence et la denoncer franchement. We have to stop taking the gloves off with this violence and denounce it frankly.

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.

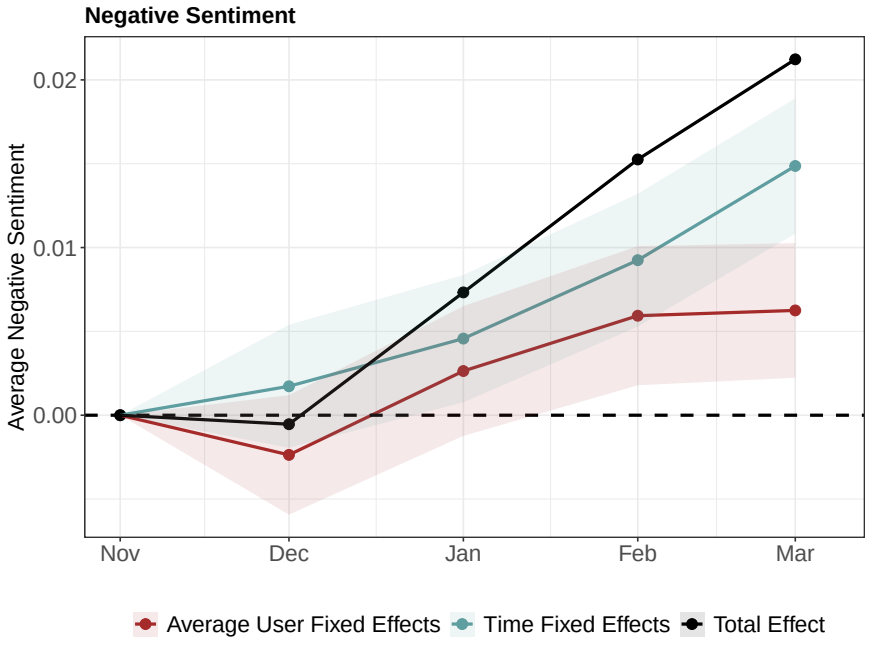
To overcome these problems, we take advantage of the fact that users can react to Facebook posts, using the following reactions: *love*, *haha*, *wow*, *angry*, *sad*. For each post in our corpus, we compute the weekly share of each of these reactions, displayed in Figure E.4. The share of *angry* reactions 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

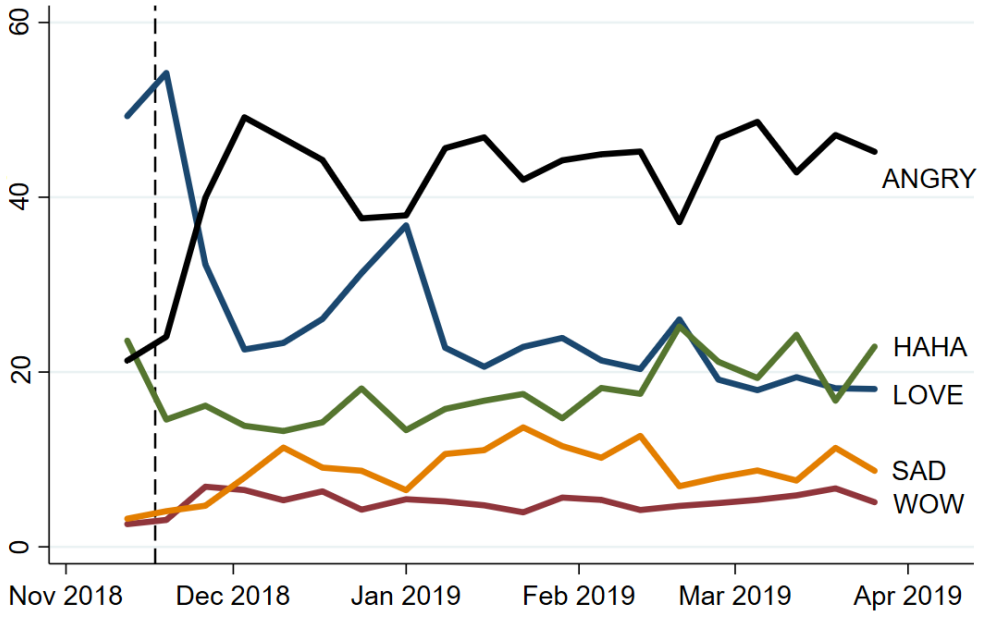
Notes: Sentences can be long and with many repetitions. For readability, we remove sequences of repeated tokens. The original phrases in French are in italics. Their English translation follows.

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

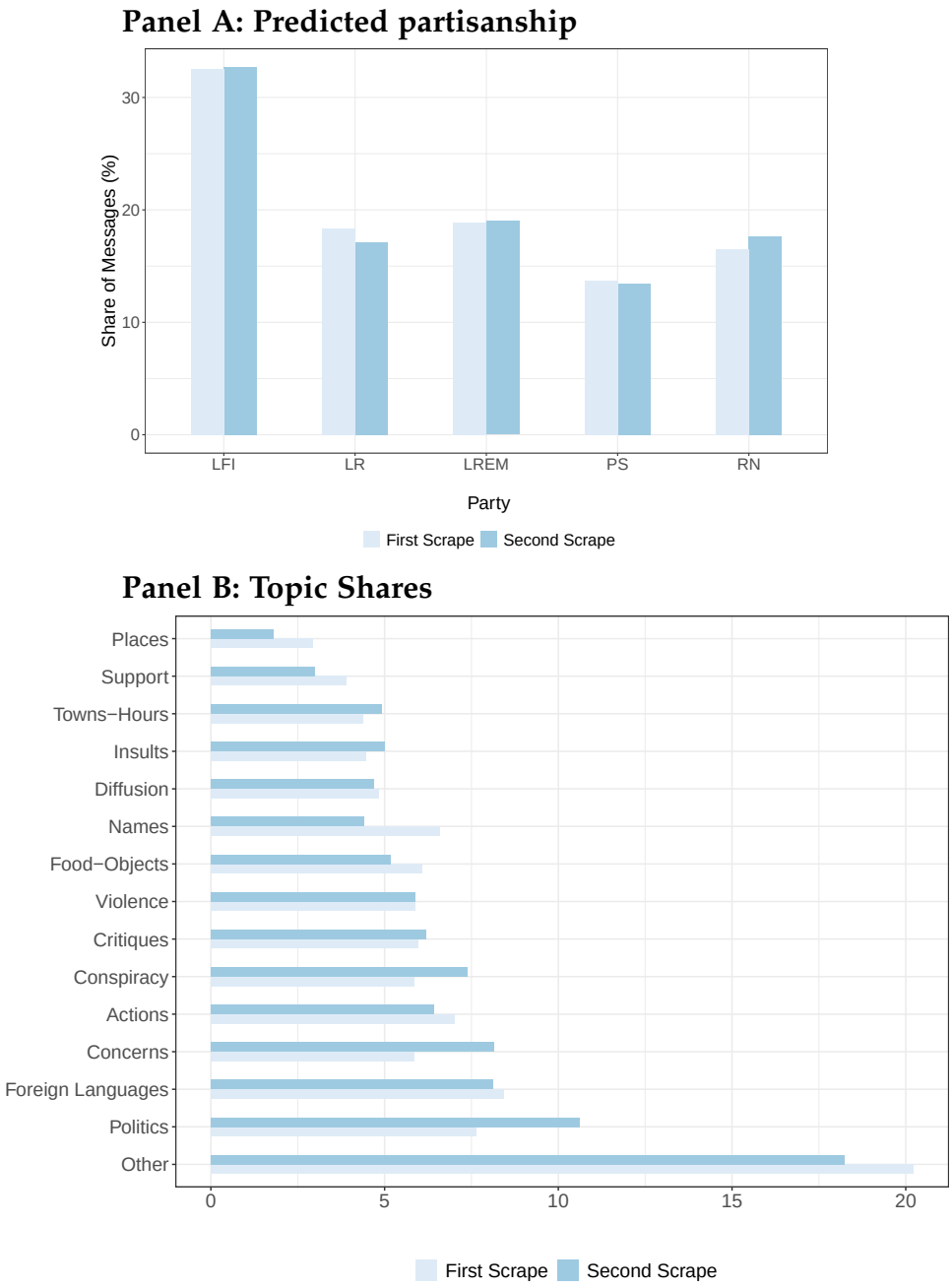
Notes: The confusion matrix C is such that C_{ij} is equal to the share of observations known to be of party i and predicted to be of party j.

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

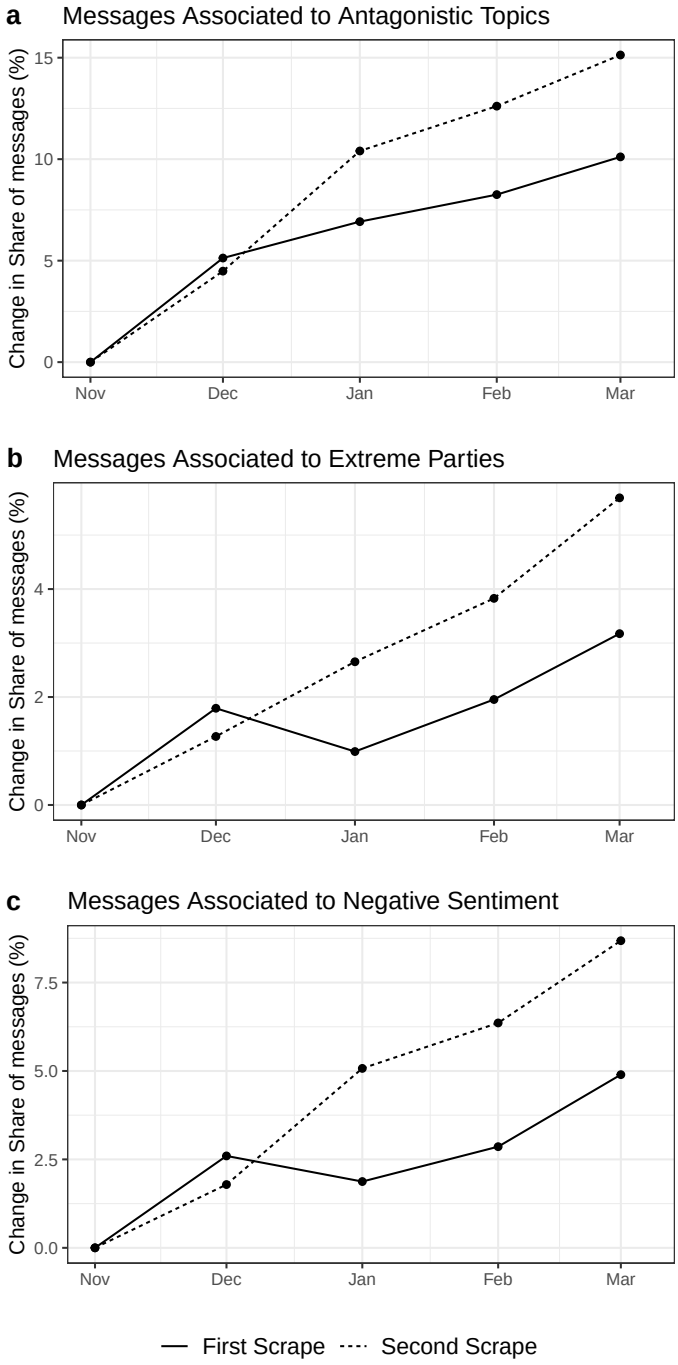
Notes: This table lists each party's 30 most predictive words according to our classifier. Words with large positive coefficients are most predictive of the speaker's party, so we simply rank the coefficients of words in descending order for each party to identify the top features.

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.

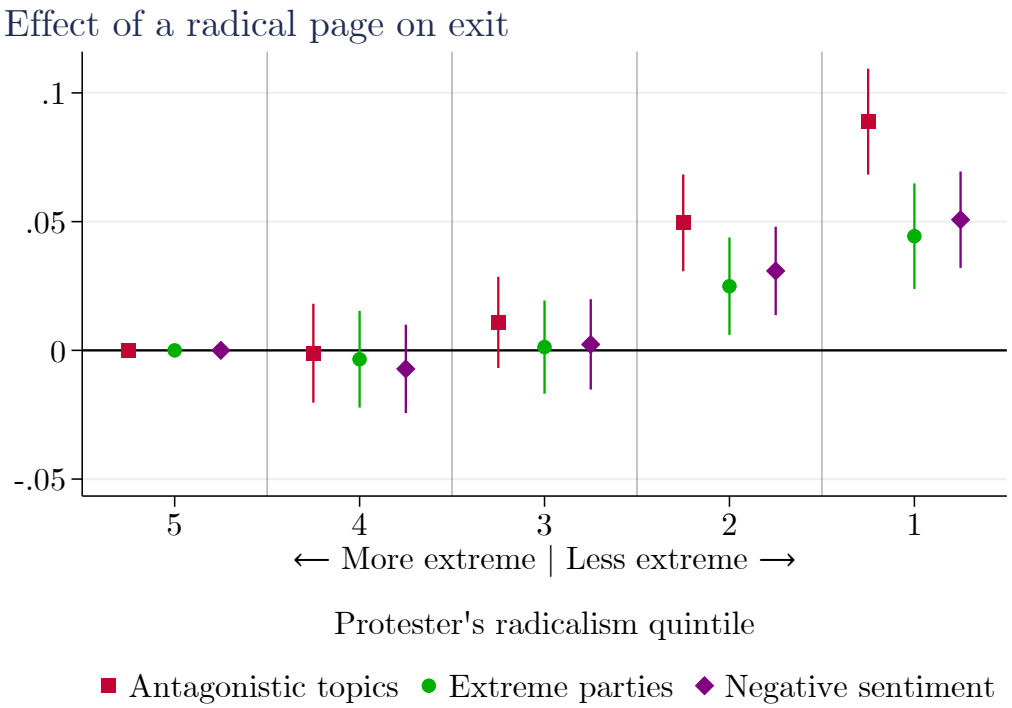
Figure E.6: Comparison of the Trends in Radical Attitudes for Each Data Collection



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.

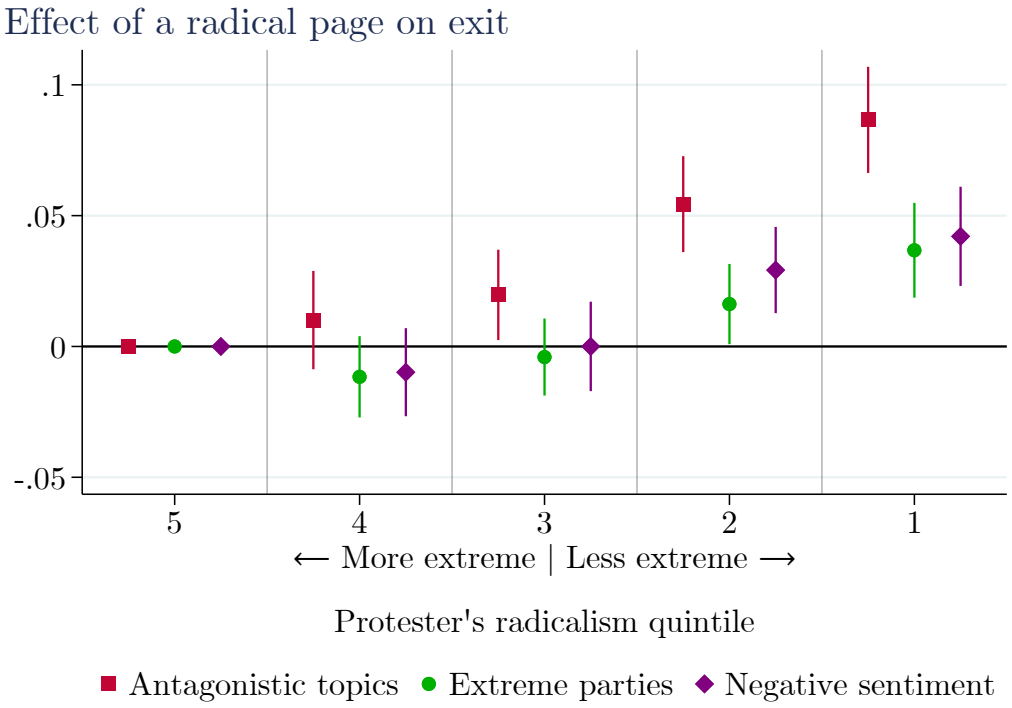
E.5 The crowd-out of moderate discussants: robustness

Figure E.7: Crowding-out over the distribution of protesters' radicalism: gross measure of page-level radicalism



Notes: This figure replicates Figure 10 using 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 6.

Figure E.8: Crowding-out over the distribution of protesters' radicalism: leave-one-out measure of page-level radicalism



Notes: This figure replicates Figure 10 using the leave-one-out average of discussants' radicalism fixed effect associated with each sentence $\mathbb{E}_{p,t,j\neq i}[\delta]$ in Equation 6 instead of the average.

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

	Probability of leaving the page				
	(1)	(2)	(3)	(4)	(5)
Panel A: Antagonistic topics					
Moderate protester × Radical page	0.023***	0.023***	0.025***	0.049***	0.039***
	(0.003)	(0.003)	(0.003)	(0.006)	(0.007)
Panel B: Extreme parties					
Moderate protester × Radical page	0.030***	0.026***	0.024***	0.030***	0.036***
	(0.003)	(0.003)	(0.003)	(0.006)	(0.007)
Panel C: Negative sentiment					
Moderate protester × Radical page	0.019***	0.026***	0.029***	0.035***	0.034***
	(0.003)	(0.003)	(0.003)	(0.006)	(0.007)
Month FE	✓	✓			
Page FE		✓			
Page-by-Month FE			✓	✓	✓
Discussant FE				✓	
Discussant-by-Month FE					✓
Observations	101,941	101,923	101,800	67,957	30,629
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

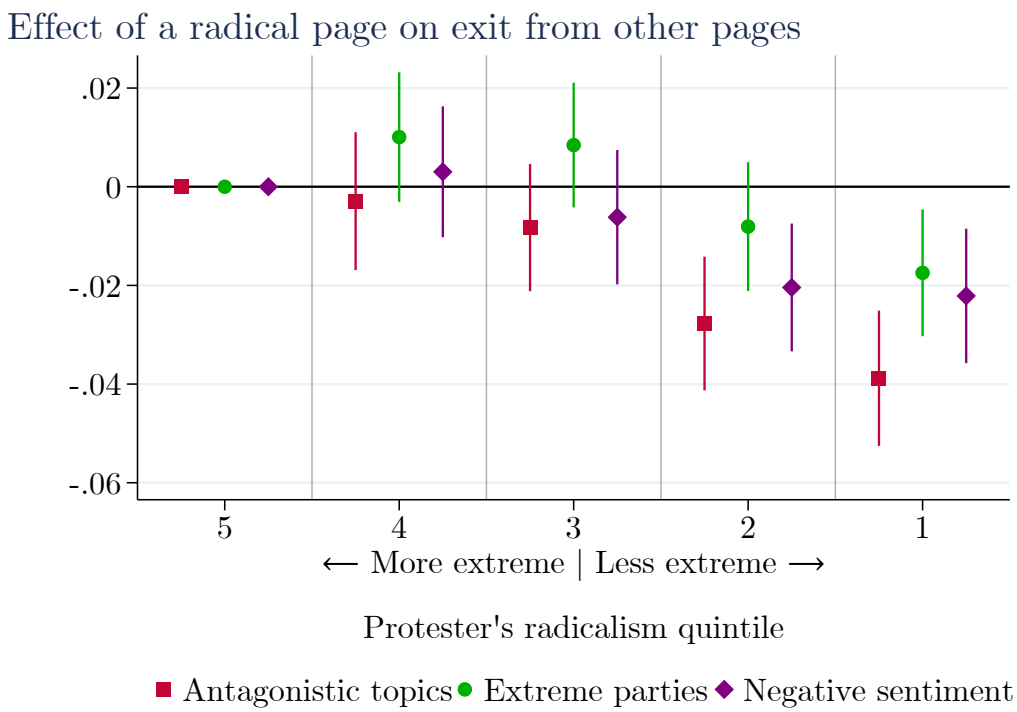
Notes: This table shows the OLS estimates of a regression of the probability of stopping posting on a Facebook page as a function of the interaction between the moderate dummy (having a fixed effect below the median of the distribution among discussants) and the (standardized) average discussant composition of the page measured at the sentence level for a given month. We replicate this exercise for three metrics of radicalization: the probability of posting a sentence associated with an antagonistic topic (Panel A), the probability of posting a sentence associated with a politically extreme party (Panel B), and negative sentiment (Panel C). 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. *: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$.

Table E.8: Local radicalization and the departure of the moderates: alternative specifications with fixed sample

	Probability of leaving the page				
	(1)	(2)	(3)	(4)	(5)
Panel A: Antagonistic topics					
Moderate protester × Radical page	0.010	0.017***	0.018***	0.045***	0.039***
	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)
Panel B: Extreme parties					
Moderate protester × Radical page	0.022***	0.022***	0.023***	0.034***	0.036***
	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)
Panel C: Negative sentiment					
Moderate protester × Radical page	0.019***	0.026***	0.027***	0.036***	0.034***
	(0.006)	(0.005)	(0.005)	(0.007)	(0.007)
Month FE	✓	✓			
Page FE		✓			
Page-by-Month FE			✓	✓	✓
Discussant FE				✓	
Discussant-by-Month FE					✓
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

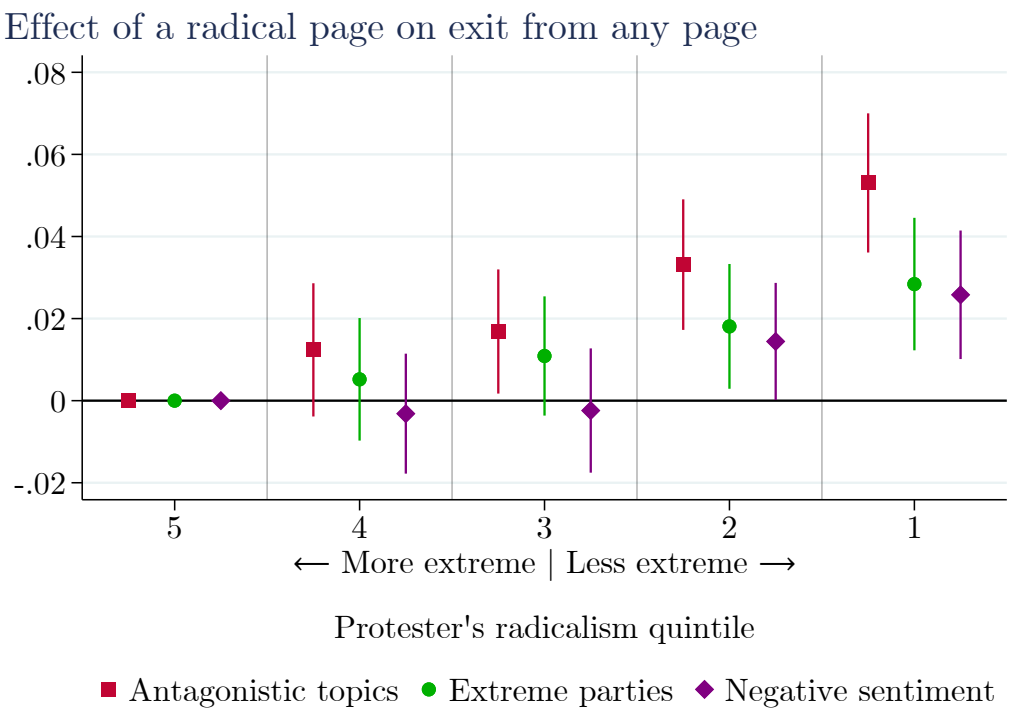
Notes: This table replicates the results shown in Table E.7 on the most restrictive sample of Column (5). We cluster standard errors at the discussant level. *: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$.

Figure E.9: Crowding-out over the distribution of protesters' radicalism: spillovers



Notes: This figure replicates Figure 10 using the probability to leave any other page the next month as the outcome variable.

Figure E.10: Crowding-out over the distribution of protesters' radicalism: overall exit



Notes: This figure replicates Figure 10 using the probability to leave any page the next month as the outcome variable.

E.6 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.9. 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. See Matter and Hodler (2024) on the impact of web search personalization.

Table E.9: Comments' Rank and Radical Content

	Measure of comments' prominence			
	(1)	(2)	(3)	(4)
Panel A: Rank of the comment (in log)				
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: Comment is among the first four (in %)				
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.

Matter, Ulrich and Roland Hodler, "Web Search Personalization During the US 2020 Election," *CEPR Discussion Paper*, 2024, (18908).

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.

Stock, James H. and Motohiro Yogo, "Testing for Weak Instruments in Linear IV Regression," in Donald W.K. Andrews, ed., *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, Cambridge University Press, 2005, pp. 80–108.

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.