Legitimize through Endorsement*

Andrea Gallice† Edoardo Grillo‡

August 17, 2022

Abstract

Individuals differ in their propensity to violate social norms. Over time, the propensity of some individuals to violate these norms changes in response to socioeconomic shocks. When these changes are not publicly observable, norm abidance may remain high because individuals fear social costs. We study how an opinion leader who is privately informed about the direction and size of the societal change can boost or hinder the abidance by a social norm. We show that the opinion leader can impact individuals' behavior when she is neither too ideologically sided in favor of the norm violation, nor too concerned about her own popularity. The impact of the opinion leader is stronger when social concerns are an important driver of individuals' behavior, the uncertainty concerning the depth of the societal change is high, and citizens interact more often with like-minded individuals.

JEL Classification: C72, D72, D83.

Keywords: Social norms; societal change; opinion leaders; endorsements; legitimization.

*We thank seminar participants at the 2022 North American Summer Meeting of the Econometric Society (Miami) and Collegio Carlo Alberto for valuable suggestions.

†ESOMAS Department, University of Torino, Corso Unione Sovietica 218bis, 10134, Torino, Italy and Collegio Carlo Alberto, Piazza Arbarello 8, 10122, Torino, Italy. *Email*: andrea.gallice@unito.it

‡Department of Economics and Management "Marco Fanno", University of Padova, Via del Santo 22, 35123, Padova, Italy. *Email*: edoardo.grillo@unipd.it

1 Introduction

In 2012, while campaigning to win a second term as President of the United States, Barack Obama publicly declared: "[...] I think same-sex couples should be able to get married." Obama's previous stance on this issue had been more nuanced: he had supported civil unions, while opposing same-sex marriages. The statement got vast media coverage. Remarkably, the net approval toward same-sex marriages in the US increased from +2% in early May 2012 to +7% in late November 2012, shortly after Obama won reelection.[1]

In the midst of the 2015 European migrant crisis, German Chancellor Angela Merkel famously declared "We will manage it!". Merkel aimed at convincing her public opinion (and, more generally, the European one) that integrating Syrian refugees was feasible. Although praised by many commentators, this pro-migrant declaration had little effect on the public opinion and it may have even politically backlashed. In the following regional and federal elections, the right-wing, xenophobic party "Alternative for Germany" obtained historically high electoral support.

Why did Obama's statement align with an ongoing societal change, and possibly contributed to it, while Merkel's did not? When can a leader's endorsement modify social norms? When does a leader's declaration have the largest impact on aggregate behavior?

We address these questions through a model of information transmission enriched with social pressure. In our model, individuals have heterogeneous propensities to violate an established social norm. A shock hits the society and shifts the attitudes of a minority of individuals towards the social norm. For instance, a youth generation becomes more open toward certain civil rights or environmental issues; or the impoverishment of the middle class fosters anti-migrant and xenophobic sentiments among the individuals who fall behind. Individuals in this minority (the novel group) know the extent of the change, while the remaining majority (the traditional group) is uncertain about it.

Individuals are then randomly matched to play a coordination game in which they must decide whether to abide by the current social norm or to violate it. The interpretation of what constitutes a violation of the norm is broad: it ranges from publicly stating a fringe opinion, to taking a controversial action, from breaking a religious taboo, to violating a commonly accepted routine. Norm-violating behavior can encompass both

[1] Source: Gallup polls (`https://news.gallup.com/poll/1651/gay-lesbian-rights.aspx`).

progressive behaviors (e.g., an expansion of civil rights) and regressive ones (e.g., an increase in discriminatory behaviors). Individuals who want to violate the norm face a coordination problem: if they break the norm, but their match does not, they suffer a social cost. Indeed, individuals who conform to a social norm often react with stigma and open hostility against deviant behaviors. The magnitude of this social cost is the *entrenchment of the social norm.*

Before playing the coordination game, individuals listen to an opinion leader (e.g., a political figure, a religious leader, or a widely-known pundit). The opinion leader holds some imperfect information about the shock that hit the society. We model this assuming that the opinion leader privately observes a binary signal correlated with the shock. A *positive signal* suggests that the novel group is more inclined to violate the norm compared to the traditional group. A *negative signal* suggests that the novel group is less inclined to violate the norm compared to the traditional group. After observing the signal, the opinion leader decides whether to publicly endorse the norm-violating behavior. The opinion leader faces a trade-off: she is ideologically inclined towards the violation of the norm, but the endorsement entails a popularity cost that is proportional to the share of citizens who keep abiding by the norm. Opinion leaders differ in the strength of their ideological motivation against the social norm relative to the popularity cost. We refer to this characteristic as to the opinion leader's type. Opinion leaders with high types accept large popularity costs to take a stance against the current norm; opinion leaders with low types do not. In our baseline model, individuals know the type of the opinion leader (for instance because they observed her past behaviors); we consider the case of uncertain type in an extension.

Our first result identifies which types of opinion leader can hinder or boost societal change through their endorsement decisions. In equilibrium, an opinion leader can modify the behavior of individuals if and only if her type is neither too high nor too low. An opinion leader with a high type disregards the popularity cost and behaves ideologically: she endorses the norm-violating behavior when she receives a positive signal, but also when she receives a negative signal. In equilibrium, her endorsement provides no information and individuals ignore it. On the contrary, an opinion leader with a low type prefers to avoid the expected popularity cost associated to the endorsement: she refrains from endorsing the violation of the norm when she receives a negative signal, but also when she receives a positive signal. In equilibrium, the lack of endorsement provides no information and individuals ignore it as well. The opinion leader can thus

impact societal behavior when she is neither too ideologically sided in favor of the norm violation, nor too concerned about her popularity.

Going back to our introductory examples, President Obama was the ideal testimonial to foster a change in the attitude towards same-sex marriages. His previous, more nuanced, stance on this issue gave credibility to his endorsement. At the same time, the popularity cost associated with the statement was low: given his overall popularity, such statement could not endanger his reelection. Chancellor Merkel, instead, had been accused by members of her own coalition to be excessively soft on migration.[2] Her statement then came across more as her individual wish than as a fair description of Germans' attitude towards migrants. As such, she was unable to foster a change in societal behavior.

We then study how some key features of the society affect *the scope* of the opinion leader's endorsement. These features include the size of the novel group, the uncertainty concerning the magnitude of the shock, and the entrenchment of the social norm. When the opinion leader decides whether to endorse the norm-violating behavior or not, she compares her private ideological benefit from the endorsement with the expected popularity cost. If the novel group grows larger in size, or if the uncertainty concerning the average preference in the novel group gets larger, the signal of the opinion leader becomes more informative. The expected popularity cost after a negative signal thus increases, while the expected popularity cost after a positive signal decreases. Both these changes push the opinion leader to truthfully reveal the signal she received. The set of opinion leaders who, in equilibrium, impact societal behavior then widens. Instead, if the entrenchment of the social norm increases, the expected popularity cost of an endorsement increases after both signals. Only opinion leaders with strong ideological motivation can thus impact societal behavior. These results suggest that opinion leaders are more likely to shape the behavior of societies undergoing potentially deep transformations (e.g., young societies, or societies that experience large migration flows or economic shocks). Furthermore, in societies where a social norm is deep-rooted, successful advocates against it ought to exhibit radical preferences.

We also investigate *the impact* of the opinion leader. This is defined as the share of individuals who, in equilibrium, modify their behavior in response to the opinion leader's

[2]For instance, Horst Seehofer, an historical ally of Merkel and the leader of the CSU party, openly criticized the Chancellor's migration speech declaring: "With the best will, I can't embrace this sentence." (Source: `https://apnews.com/article/0170714f16cc46f39ce03e85e825126e`)

endorsement (or lack thereof). We find that the impact is larger when the signal of the opinion leader is more informative and when the social norm is more entrenched.

We then extend our analysis in three directions. First, we introduce homophily: we allow individuals with a propensity to violate the norm (or not to violate it) to interact more often with individuals who share a similar propensity. We find that homophily increases the impact of the opinion leader. Second, we show that the insights of our model hold true even if we introduce uncertainty about the opinion leader's type. Finally, we introduce multiple opinion leaders who independently decide whether to endorse the violation of the norm. We show that the existence of multiple opinion leaders does not affect the credibility of each of them. Yet, multiple endorsement decisions can now either reinforce or offset each others. The impact of each opinion leader then depends on what all the others do.

1.1 Literature Review

In our model, the opinion leader eases or hinders societal change through her endorsement decision. We thus contribute to the literature that investigates how social norms evolve over time. The bulk of this literature focuses on the long-run evolution of social norms and highlights the role of history (Alesina et al. 2013, Acemoglu and Jackson 2015) and institutions (Benabou and Tirole 2011, Acemoglu and Jackson 2017). In contrast, we study how opinion leaders can shape individuals' behavior in the short-run. In this respect, we are close to Loeper et al. (2014), Carlsson et al. (2016), Bursztyn et al. (2020), Müller and Schwarz (2020), and Grosjean et al. (2021).

Among the papers that take a long-run perspective, Acemoglu and Jackson (2015) is the most related to us. In Acemoglu and Jackson (2015), agents play a coordination game over multiple periods. Some "prominent" individuals with greater visibility can influence the expectations, hence the behavior, of future generations. In Acemoglu and Jackson (2015) individuals differ in their exogenous prominence. In our setting, instead, the preferences of opinion leaders endogenously affect their ability to impact society.

This last feature also distinguishes us from Bursztyn et al. (2020). Like us, Bursztyn et al. (2020) consider a setting in which agents are uncertain about other individuals' preferences and abide by social norms. In their model, individuals update their beliefs about the society based on an exogenous public signal, say a surprising electoral

outcome.[3] In this paper, we study the strategic behavior of an opinion leader who provides information through her endorsement decision. This enables us to investigate the interplay between the opinion leader's preferences and societal behavior.

Our paper investigates how the scope and impact of the opinion leader's endorsement decision vary with some key features of the society. This distinguishes our work from Loeper et al. (2014) that study a model in which individuals first observe a random sample of actions taken by biased experts and then decide what to do. Furthermore, in Loeper et al. (2014), experts impact individual choices only if individuals are uncertain about the experts' bias/type. This is due to the joint effect of coordination motives and social learning. In our setting, instead, opinion leaders can affect individual and aggregate behavior even when their types are common knowledge.[4]

We study the credibility of opinion leaders' endorsements. We are thus related to the literature on strategic information transmission. In our setting, the endorsement of the norm-violating behavior entails an ideological benefit and a popularity cost. In this respect, we are close to models of information transmission with ideological biases (Cowen and Sutter 1998, Cukierman and Tommasi 1998) and reputational concerns (Morris 2001, Ottaviani and Sørensen 2006a, Ottaviani and Sørensen 2006b). This last feature also links our paper to the literature on pandering (Che et al. 2013, Morelli and Van Weelden 2013, Gratton 2014, and Maskin and Tirole 2019). Within the literature on information transmission, our work is also related to papers that study the persuasion by experts or politicians (see, among others, Jackson and Tan 2013, Schnakenberg 2015, 2017, Alonso and Câmara 2016, Chan et al. 2019, Gulotty and Luo 2021, Gerardi et al. 2022, Prato and Turner 2022). We contribute to this literature studying how the ideology of the opinion leader and the norm entrenchment affect her ability to shape societal behavior. The focus on the norm entrenchment also distinguishes us from the literature on media bias (Mullainathan and Shleifer 2005, Baron 2006, Gentzkow and Shapiro 2006, Prat and Strömberg 2013).

[3] Bursztyn et al. (2020) provide evidence that Trump's victory in the 2016 US presidential election increased individuals' willingness to express xenophobic views and made such opinions more socially acceptable. In a similar vein, Müller and Schwarz (2020) show that Trump's tweets concerning Islam-related topics triggered anti-Muslim hate crimes, whereas Grosjean et al. (2021) find evidence that Trump's rallies boosted racial prejudice against minorities.

[4] In Carlsson et al. (2016), opinion leaders are heterogeneous, but this heterogeneity is in terms of quality rather than ideology/office motivation. Exploiting this quality heterogeneity, Carlsson et al. (2016) explain why, once in power, some politicians generate consensus on debated issues, while others do not.

Insofar we study the role of opinion leaders in shaping individual behavior, our work is related to a recent literature that investigates the market for online endorsements (Fainmesser and Galeotti 2021, Hinnosaar and Hinnosaar 2021, Mitchell 2021). In these models, firms hire influencers to advertise products. The opinion leader in our model does not receive a direct compensation out of her endorsement and there is no third party trying to buy her support.

Finally, in our model, individuals face a coordination problem in the presence of a social norm. Violating this norm entails a social cost. We are thus related to papers that highlight the relevance of social pressure for individual and collective choices (see, for instance, Bernheim 1994, Hopkins and Kornienko 2004, Levy and Razin 2015, Gallice and Grillo 2020, Friedrichsen et al. 2021 and the references therein).

2 The Model

A society is made by a unit mass of individuals ("he") and by an opinion leader ("she"). Individuals interact for two consecutive periods, $t = 1, 2$.

In period 1, each individual decides independently and simultaneously whether to abide by a prevailing social norm, action $a_i = 0$, or to violate it, action $a_i = 1$. We refer to individuals choosing $a_i = 0$ as *abiders* and to individuals choosing $a_i = 1$ as *violators*.

After individuals have chosen their actions, they are randomly matched in pairs and each individual gets a payoff determined according to Table 1.

i \ j	$a_j = 0$	$a_j = 1$
$a_i = 0$	$0, 0$	$0, \theta_j - \lambda$
$a_i = 1$	$\theta_i - \lambda, 0$	θ_i, θ_j

Table 1: Payoffs from social interaction.

Individual i thus enjoys a safe payoff equal to zero when he abides by the social norm. Instead, he enjoys an hedonic *private payoff* equal to $\theta_i \in \mathbb{R}$ when he violates the norm. Private payoffs are i.i.d. in the population and their distribution is uniform in the interval $[-\gamma, \gamma]$.[5] The parameter $\gamma \in \mathbb{R}_{++}$ measures the heterogeneity of individuals' private payoffs; we refer to it as to the *baseline heterogeneity*.

[5] The uniform distribution provides analytic tractability, but our results immediately extend to other distributions.

The social norm is entrenched: an individual suffers a social cost equal to λ if he chooses to violate the norm, but his match abides by it. Parameter $\lambda \in \mathbb{R}_+$ measures the *norm entrenchment*.

In our baseline model, all matches are equally likely (see Section 5.1 for the case in which individuals are more likely to encounter like-minded individuals). The expected payoff of an individual with private payoff θ_i is thus equal to

$$u(a_i, \overline{a}_1; \theta_i) = a_i[\theta_i - (1 - \overline{a}_1)\,\lambda], \tag{1}$$

where $\overline{a}_1$ is the share of violators in period 1.

At the end of period 1, the preferences of a share α of the population change. Some individuals may die and be replaced by new ones with different preferences; or some societal change (e.g., immigration, diffusion of new ideas, economic shocks, and so on) may modify individuals' attitude towards the prevailing social norm. Social changes are gradual processes. We capture this assuming that the shock modifies the preferences of a minority of the population: $\alpha \in (0, 1/2)$.

In period 2, the population thus consists of two groups. A share $(1-\alpha)$ of individuals belongs to the *traditional group*. These individuals have the same private payoffs as of period 1. Instead, a share α belongs to the *novel group*. These individuals have private payoffs that are i.i.d. draws from a uniform distribution in the interval $[\omega - \gamma, \omega + \gamma]$. The average difference in the propensity to violate the social norm among the two groups is measured by ω, the *depth of the societal change*. The value of ω is private information of the individuals in the novel group. Individuals in the traditional group believe that ω is uniformly distributed in the interval $[-\psi, \psi]$. The parameter $\psi \in \mathbb{R}_{++}$ measures the *uncertainty of the societal change*.

The societal change is thus captured by the pair (α, ω). The parameter α captures the extensive margin of the change: it measures the share of the population with novel preferences. The parameter ω captures the intensive margin: it measures the average difference in the attitude towards the social norm between the two groups.

In period 2, individuals are again randomly matched and they play the coordination game summarized in Table 1. The match in period 2 is independent of the match in period 1. Thus, the expected payoff of an individual with private payoff θ_i in the second period is

$$u(a_i, \overline{a}_2; \theta_i) = a_i[\theta_i - (1 - \overline{a}_2)\,\lambda], \tag{2}$$

where $\overline{a}_2$ is the share of violators in period 2.

Differently from period 1, though, before individuals choose their actions, an opinion leader can endorse the violation of the social norm ($b = 1$) or not ($b = 0$). The opinion leader's endorsement (or lack thereof) becomes common knowledge as soon as it occurs. The opinion leader has private information concerning the depth of the societal change, ω. By virtue of her role, her daily interactions with people, or her preferential access to public opinion polls, the opinion leader becomes aware of in-progress societal changes. In particular, the opinion leader observes a private signal $s \in \{0, 1\}$ where:

$$\Pr(s = 0 \mid \omega) = \frac{1}{2} - \frac{\omega}{2\psi} \quad \text{and} \quad \Pr(s = 1 \mid \omega) = \frac{1}{2} + \frac{\omega}{2\psi}.$$

The likelihood of signal $s = 0$ is thus higher than the likelihood of signal $s = 1$ when ω is negative. The opposite is true when ω is positive. We refer to $s = 0$ as to the negative signal and to $s = 1$ as to the positive signal. A positive (negative) signal suggests that the novel group is on average more (less) inclined to violate the norm than the traditional group.

When the opinion leader chooses not to endorse the violation of the norm, $b = 0$, she gets a payoff equal to 0. When the opinion leader endorses the violation of the norm, $b = 1$, she gets a private payoff equal $k \in (0, 1)$, but she also experiences a popularity cost proportional to the share of individuals who keep abiding by the social norm. The weight the opinion leader puts on the popularity cost is equal to $(1 - k)$. The payoff of the opinion leader is thus equal to:

$$v(b; \overline{a}_2) = b\left[k - (1 - k)(1 - \overline{a}_2)\right]. \tag{3}$$

At the end of period 2 the game ends.[6] We define $K \equiv k/(1 - k) \in (0, \infty)$ as *the ideological strength* of the opinion leader. It measures the relative strength of the opinion leader's private payoff over her popularity concerns. In our baseline model, K is common knowledge. Individuals observe the past record of the opinion leader and identify her

[6] Our analysis immediately extends to infinite-horizon settings in which either of the following two conditions hold. First, at end of each period t, individuals observe the average action in the society $\overline{a}_t$, but at the beginning of period $t+1$ a new shock hits the society and modifies the preferences of another group of individuals. Second, at the end of each period t, individuals only observe the action taken by their match and have no feedback on the average action chosen in the society, $\overline{a}_t$. If we allow for more general feedback concerning the aggregate behavior in period $t-1$, the opinion leader could still ease or hinder societal changes as long as individuals cannot perfectly forecast the aggregate behavior at time t, $\overline{a}_t$.

attitude towards the social norm. Section 5.2 shows that the insights of the paper hold true even when individuals' are uncertain about the opinion leader's ideological strength.

Finally, we assume that the baseline heterogeneity is large enough to guarantee that some individuals always violate (respectively, abide by) the social norm no matter what others do.

Assumption 1. *In both periods, some individuals always violate the social norm, while others always abide by it:* $\gamma \geqslant \max\{\lambda, \psi\}$.

Assumption 1 guarantees that the opinion leader can always impact societal behavior.

We solve the game using perfect Bayesian equilibrium.[7] We refer to this solution concept simply as to the equilibrium of the game.

3 Equilibrium Analysis

In period 1, individuals optimally follow a cutoff strategy: an individual violates the norm if and only if his private payoff θ_i exceeds a cutoff $\overline{\theta}_1$.[8] Under this cutoff strategy, individuals with private payoffs above (below) $\overline{\theta}_1$ are violators (abiders). An individual with private payoff equal to $\overline{\theta}_1$ must thus be indifferent between abiding by the norm or violating it: $\overline{\theta}_1 = \left(1 - \int_{\overline{\theta}_1}^{\gamma} \frac{dx}{2\gamma}\right)\lambda$. This indifference conditions yields $\overline{\theta}_1 = \lambda\gamma/(2\gamma - \lambda)$ and the share of violators in period 1 is then equal to

$$\overline{a}_1 = \frac{\gamma - \lambda}{2\gamma - \lambda}. \tag{4}$$

The share of violators is decreasing in the norm entrenchment, λ, and increasing in the baseline heterogeneity, γ. When λ is large, the expected cost of violating the norm goes up and only individuals with high private payoffs remain violators. Instead, when γ is large, more individuals exhibit extreme private payoffs and violate the social norm independently of the expected social cost. The share of violators thus increases.

In period 2, only individuals in the novel group know the realization of ω. Individuals in the traditional group do not. We can thus define two cutoff strategies:[9] a state-

[7] We restrict attention to equilibria in which the opinion leader can only choose whether to endorse or not the violation of the norm. Given the signal space, this is without loss of generality. For instance, after a message that has probability zero, we can assume that individuals believe that the opinion leader sends such message with the same probability independently of the signal she received.

[8] The utility function of individuals defined in equation (1) satisfies the single-crossing property in θ_i. The specific tie-breaking rule when $\theta_i = \overline{\theta}_1$ does not affect our analysis.

[9] The optimality of cutoffs strategy follows from the same reason highlighted in footnote 8

independent cutoff strategy with threshold $\overline{\theta}_2^T$ for the traditional group and a state-dependent cutoff strategy with threshold $\overline{\theta}_2^N(\omega)$ for the novel group. When individuals follow these cutoff strategies, an individual in group $j \in \{T, N\}$ with private payoff equal to $\overline{\theta}_2^j$ is indifferent between abiding by the norm or violating it. Let $\overline{a}_2^T$ and $\overline{a}_2^N(\omega)$ be the share of violators in the traditional and in the novel group (see the proof of Proposition 1 for details). The overall share of violators in period 2 is equal to $\overline{a}_2(\omega) = (1-\alpha)\overline{a}_2^T + \alpha\overline{a}_2^N(\omega)$.

Proposition 1. *In equilibrium, the share of violators in period* 2 *is equal to:*

$$\overline{a}_2(\omega) = \overline{a}_1 + \frac{\alpha}{2\gamma - \alpha\lambda}\left(\omega + \frac{(1-\alpha)\lambda}{2\gamma - \lambda}\mathbb{E}\left[\omega \mid \mathcal{I}_2^T\right]\right),$$

where $\mathcal{I}_2^T$ is the information available to individuals in the traditional group when they choose their action in period 2.

The share of violators in the second period differs from the one in the first period in two respects. First, the shock ω shifts the preferences of an α-share of the population. Second, individuals in the traditional group form an expectation about ω and react to it. This expectation directly impacts the behavior of individuals in the traditional group, but it also indirectly affects the behavior of individuals in the novel group. Although individuals in the novel group know the value of ω, they care about social payoffs and they thus react to $\mathbb{E}\left[\omega \mid \mathcal{I}_2^T\right]$ as well.

The endorsement decision of the opinion leader affects the share of violators through its impact on $\mathbb{E}\left[\omega \mid \mathcal{I}_2^T\right]$. To understand the impact of the opinion leader on the share of violators, we first characterize the benchmark case in which the opinion leader does not exist. In this case, $\mathbb{E}[\omega \mid \mathcal{I}_2^T] = 0$.

Remark 1. *Suppose the opinion leader does not exist. The share of violators in period* 2 *is then equal to*

$$\overline{a}_2^{NL}(\omega) = \overline{a}_1 + \frac{\alpha}{2\gamma - \alpha\lambda}\omega.$$

Individuals in the traditional group behave as in period 1*, while individuals in the novel group adjust their behavior in response to the realization of ω.*

When there is no opinion leader, the share of violators is higher (lower) than in period 1 if and only if the novel group is on average more inclined to violate the social norm compared to the traditional group ($\omega > 0$).

3.1 Informative Equilibria

Consider now the case in which the opinion leader exists. We first focus on *informative equilibria*; these are equilibria in which the beliefs of individuals react to the endorsement decision of the opinion leader.

We can summarize the behavior of the opinion leader with an endorsement strategy, namely a pair $(\beta(0), \beta(1)) \in [0,1]^2$ where $\beta(s)$ is the probability with which the opinion leader endorses the violation of the norm after signal $s \in \{0,1\}$. The endorsement strategy is informative if $\beta(0) \neq \beta(1)$. The signal structure implies that, holding her behavior constant, the opinion leader believes that, in equilibrium, fewer individuals violate the norm after signal $s = 0$ than after signal $s = 1$. Hence, we focus on endorsement strategies in which $\beta(0) < \beta(1)$: the opinion leader endorses the violation of the norm less often after signal $s = 0$ than after signal $s = 1$.

The endorsement strategy is *fully informative* if $(\beta(0), \beta(1)) = (0,1)$. In this case, individuals perfectly infer the opinion leader's signal from her endorsement decision. Instead, when the endorsement strategy satisfies $0 \leqslant \beta(0) < \beta(1) \leqslant 1$ with at least one of the two weak inequalities being strict, the endorsement strategy is *partially informative*. In this case, individuals update their beliefs based on the opinion leader's endorsement decision, but they do not always infer the signal she received.

In an informative equilibrium, the expectation about ω held by individuals in the traditional group are:

$$\mathbb{E}[\omega \mid b] = \begin{cases} -\frac{[\beta(1)-\beta(0)]}{[2-\beta(1)-\beta(0)]} \cdot \frac{\psi}{3} & \text{if } b = 0 \\ \frac{[\beta(1)-\beta(0)]}{[\beta(1)+\beta(0)]} \cdot \frac{\psi}{3} & \text{if } b = 1. \end{cases}$$

Proposition 1 then implies that in a fully informative equilibrium ($\beta(0) = 1$ and $\beta(1) = 1$) the share of violators in period 2 is equal to:

$$\overline{a}_2^{FI}(\omega \mid b = 0) = \overline{a}_1 + \frac{\alpha}{2\gamma - \alpha\lambda}\left(\omega - \frac{(1-\alpha)\lambda}{2\gamma - \lambda} \cdot \frac{\psi}{3}\right), \tag{5}$$

$$\overline{a}_2^{FI}(\omega \mid b = 1) = \overline{a}_1 + \frac{\alpha}{2\gamma - \alpha\lambda}\left(\omega + \frac{(1-\alpha)\lambda}{2\gamma - \lambda} \cdot \frac{\psi}{3}\right). \tag{6}$$

In a fully informative equilibrium, the endorsement decision of the opinion leader fully reveals the information available to her. When the opinion leader endorses the violation

of the norm, the share of violators is higher than in Remark 1. The opposite is true when the opinion leader does not endorse the violation.

The payoff of the opinion leader depends on the share of violators and this share changes with her endorsement decision. Thus, her endorsement decision does not necessarily reflect the signal she received. An opinion leader with high ideological strength ($K \equiv k/(1-k)$ large) endorses the violation of the norm even when she receives signal $s = 0$. Her ideological motivation is so strong that she takes a stance against the norm even when the expected popularity cost is high. An opinion leader with strong popularity concerns (K low), instead, never endorses the violation of the norm. Because the signal she receives is noisy, the opinion leader incurs an expected popularity cost whenever she endorses the violation. If popularity concerns are strong, she prefers to avoid this cost. A fully informative equilibrium thus exists if and only if the type of the opinion leader takes an intermediate value.

Proposition 2. *A fully informative equilibrium exists if and only if* $K \in \left[\underline{K}, \overline{K}\right]$ *where*

$$\underline{K} = \frac{1}{2\gamma - \lambda}\left(\gamma - \frac{\alpha\psi}{3}\right) \quad \text{and} \quad \overline{K} = \frac{1}{2\gamma - \lambda}\left(\gamma + \frac{2\gamma - \lambda - (1-\alpha)\lambda}{2\gamma - \alpha\lambda} \cdot \frac{\alpha\psi}{3}\right).$$

In a fully informative equilibrium, the shares of violators are given by equation (5) *and equation* (6).

When the ideological strength of the opinion leader lies below $\underline{K}$, full information transmission is not credible: popularity concerns refrain her from endorsing the violation after signal $s = 1$. When the ideological strength lies above $K > \overline{K}$, full information transmission is not credible either: ideological motivation pushes the opinion leader to endorse the violation of the norm even after signal $s = 0$. Nonetheless, partially informative equilibria exist if K lies above $\overline{K}$. In these equilibria, the opinion leader endorses the violation of the norm with certainty after signal $s = 1$, but also, with some positive probability, after signal $s = 0$; that is, $\beta(1) = 1$ and $\beta(0) \in (0, 1)$.[10]

Proposition 3. *A partially informative equilibrium in which* $\beta(0) \in (0,1)$ *and* $\beta(1) = 1$ *exists if and only if* $K \in \left(\overline{K}, K^{\dagger}\right)$ *where*

$$K^{\dagger} = \frac{1}{2\gamma - \lambda}\left(\gamma + \frac{2\gamma - \lambda}{2\gamma - \alpha\lambda} \cdot \frac{\alpha\psi}{3}\right).$$

[10] A partially informative equilibrium in which $\beta(0) = 0$ and $\beta(1) \in (0,1)$ is possible only in the non-generic case in which $K = \underline{K}$ (see the proof of Proposition 3 for details). We will ignore this non-generic case.

In this partially informative equilibrium, $\beta(0)$ is increasing in K.

The equilibrium share of violators in a partially informative equilibrium is linear in K:

$$\overline{a}_2^{PI}(\omega \mid b=0) = K - \frac{\lambda}{2\gamma-\lambda} + \frac{\alpha}{2\gamma-\alpha\lambda}\left(\omega - \tfrac{\psi}{3}\right) \tag{7}$$

$$\overline{a}_2^{PI}(\omega \mid b=1) = 1 - K + \frac{\alpha}{2\gamma-\alpha\lambda}\left(\omega + \tfrac{\psi}{3}\right). \tag{8}$$

When $K = K^\dagger$, the two previous expressions are equal and correspond to $\overline{a}^{NL}(\omega)$: when K is greater or equal than $K^\dagger$, the endorsement decision of the opinion leader does not convey any information.

3.2 Uninformative Equilibria

As highlighted at the end of the previous section, *uninformative equilibria* also exist. In these equilibria, individuals do not update their beliefs based on the opinion leader's endorsement decision; that is, $\mathbb{E}[\omega \mid b=0] = \mathbb{E}[\omega \mid b=1] = \mathbb{E}[\omega] = 0$. The share of violators is thus independent of the opinion leader's behavior.

Proposition 4. *An uninformative equilibrium exists if and only if $K \notin \left[K^U, K^\dagger\right]$, where*

$$K^U = \frac{1}{2\gamma-\lambda}\left(\gamma - \frac{2\gamma-\lambda}{2\gamma-\alpha\lambda}\cdot\frac{\alpha\psi}{3}\right) \in (\underline{K}, \overline{K}).$$

In an uninformative equilibria, $\overline{a}_2^U(\omega \mid b=0) = \overline{a}_2^U(\omega \mid b=1) = \overline{a}_2^{NL}(\omega)$.

Unlike in cheap talk models, uninformative equilibria do not always exist. In our setting, endorsements carry both an ideological private benefit and an expected popularity cost. The opinion leader updates her belief about the popularity cost based on the signal she receives. If the signal is $s=1$, she believes that the cost is small. If the signal is $s=0$, she believes that the cost is large. When the ideological private benefit and the expected popularity cost are both important ($K \in [K^U, K^\dagger]$), the opinion leader adjusts her endorsement decision based on the signal and uninformative equilibria do not exist.

Figure 1 summarizes the opinion leader's behavior in the equilibria described above. It depicts $\beta(0)$ (solid red line) and $\beta(1)$ (dashed blue line) in the most informative equilibrium for different values of K. Outside the ranges identified by Propositions 2 and 3, only uninformative equilibria exist. When K is below $\underline{K}$, popularity concerns

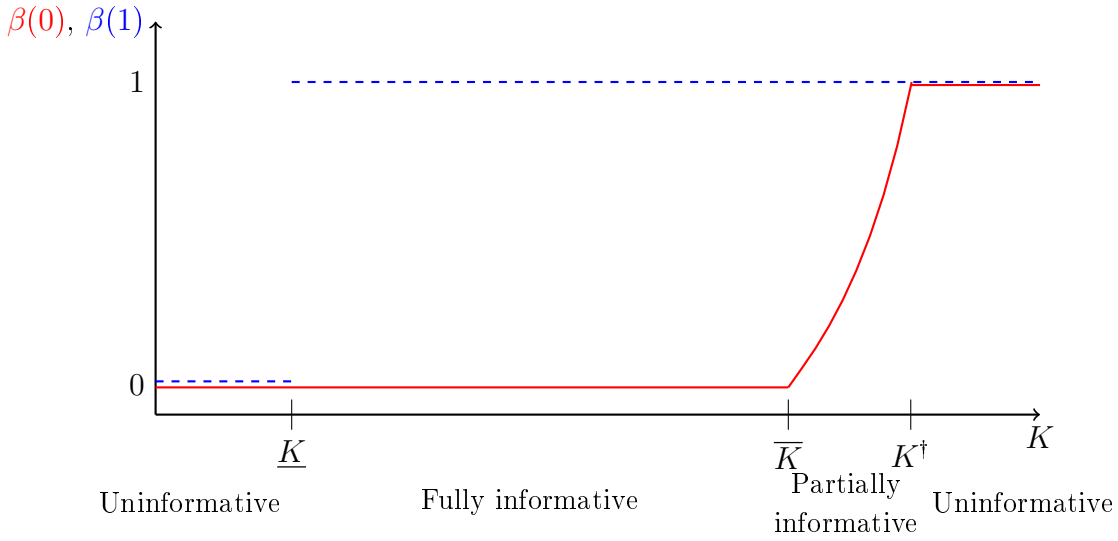

Figure 1: The opinion leader's equilibrium behavior.

Notes: The solid red line shows the probability the opinion leader endorses the violation of the norm after signal $s = 0$, $\beta(0)$. The dashed blue line shows the probability the opinion leader endorses the violation of the norm after signal $s = 1$, $\beta(1)$.

refrain the opinion leader from endorsing the violation of the norm after signal $s = 1$. In this case, $\beta(0) = \beta(1) = 0$. On the contrary, when K exceeds $K^\dagger$, the opinion leader endorses the violation of the norm even after signal $s = 0$. In this case, $\beta(0) = \beta(1) = 1$. When K takes values in the interval $\left[\underline{K}, \overline{K}\right]$, the opinion leader endorses the violation of the norm after signal $s = 1$ and she does not endorse it after signal $s = 0$. Finally, when K lies in the interval $\left(\overline{K}, K^\dagger\right)$, the opinion leader always endorses the violation of the norm after signal $s = 1$, $\beta(1) = 1$, and also with positive probability after signal $s = 0$, $\beta(0) \in (0, 1)$.

4 The Impact of the Opinion Leader

The opinion leader affects the share of violators when her ideological strength lies in an intermediate range; that is, it is neither too high, nor too low. The next proposition studies how the bounds of this range change with some key features of the society (see also Figure 1).

Proposition 5. *The range* $\left[\underline{K}, \overline{K}\right]$ *for which a fully informative equilibrium exists, widens when* α *or* ψ *increase, and shifts to the right when* λ *increases. As* γ *increases, the range shrinks if* $\frac{2}{3}\alpha\psi \geqslant \lambda$*, and moves to the left otherwise. The range* $\left(\overline{K}, K^\dagger\right)$

for which a partially informative equilibrium exists, shifts to the right when α, ψ or λ increase and it shifts to the left when γ increases.

As the size of the novel group α grows larger, the signal the opinion leader receives conveys information on a larger share of the population. Hence, the expected popularity cost associated to the endorsement goes up after signal $s = 0$, and it goes down after signal $s = 1$. This strengthens the incentives of the opinion leader not to endorse the violation of the norm after signal $s = 0$ and to endorse it after signal $s = 1$. The range $[\underline{K}, \overline{K}]$ thus widens.

Similarly, when the uncertainty concerning the societal change ψ increases, the signal becomes more informative. The expected cost of endorsing the violation after signal $s = 0$ ($s = 1$) goes up (goes down). These effects strengthen the opinion leader's incentive to match her endorsement decision with the signal she received. The range $[\underline{K}, \overline{K}]$ widens also in this case.

An increase in the entrenchment of the norm λ shifts the range $[\underline{K}, \overline{K}]$ to the right. When the entrenchment of the social norm is stronger, more individuals abide by it. The endorsement becomes more costly for the opinion leader. A fully informative equilibrium thus requires higher levels of ideological strength; that is, the range $[\underline{K}, \overline{K}]$ shifts to the right. Figure 2 illustrates how the interval $[\underline{K}, \overline{K}]$ changes as λ and α increase.

The effect of an increase in the baseline heterogeneity γ on $[\underline{K}, \overline{K}]$ is subtler. As γ goes up, more individuals exhibit extreme private payoffs. These individuals do not respond to the endorsement decision of the opinion leader: they are either unconditional abiders or unconditional violators. Since the norm is entrenched, the share of unconditional violators increases more than the one of unconditional abiders. The expected popularity cost of endorsing the violation of the norm goes down: both $\underline{K}$ and $\overline{K}$ decrease. However, as the baseline heterogeneity increases, the opinion leader's endorsement decision affects a lower share of individuals. This weakens her incentive to endorse the violation of the norm after signal $s = 1$: the threshold $\underline{K}$ goes up. The overall impact of γ on $\underline{K}$ is thus ambiguous. The second force dominates when the informativeness of the opinion leader endorsement is already low; that is, when $\alpha\psi$ is small. In this case, the interval $\left[\underline{K}, \overline{K}\right]$ moves to the left: only opinion leaders with high ideological strength can play a fully informative equilibrium. Instead, when $\alpha\psi$ is large, full information transmission becomes harder to support: $\left[\underline{K}, \overline{K}\right]$ shrinks.

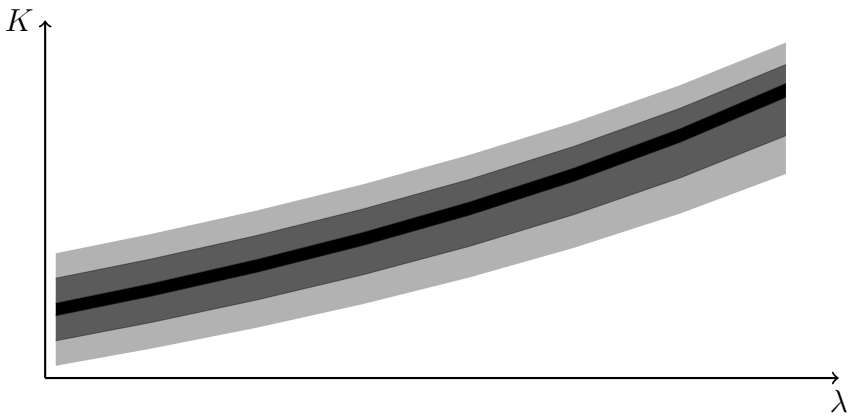

Figure 2: Range of values for which a fully informative equilibrium exists.

Notes: The figure shows the range of K for which a fully informative equilibrium exists as a function of λ when the share of the novel group is $\alpha = 0.05$ (black), $\alpha = 0.25$ (dark grey) and $\alpha = 0.45$ (light grey).

The upper bound on the range of parameters for which a partially informative equilibrium exists, $K^{\dagger}$, reacts to changes in parameters in the same way as $\overline{K}$ does.[11]

To sum up, an increase in α or ψ unambiguously reinforces the scope of the endorsement decision. An increase in λ, instead, favors information transmission among opinion leaders with high ideological strength, but impedes it among opinion leaders with low ideological strength. Finally, an increase in γ has an ambiguous effect: either it dampens the scope of the endorsement decision, or it favors it among opinion leaders with high ideological strength and impedes it among those with low ideological strength.

Next, we discuss the impact of the opinion leader's endorsement decision on the share of violators. Such impact is equal to the difference between the share of violators when the opinion leader exists and the same share when she does not exist (see Remark 1):

$$\overline{a}_2(\omega) - \overline{a}_2^{NL}(\omega) = \frac{\alpha}{2\gamma - \alpha\lambda} \cdot \frac{(1-\alpha)\lambda}{2\gamma - \lambda} \, \mathbb{E}\left[\omega \mid \mathcal{I}_2^T\right]. \tag{9}$$

Figure 3 plots the share of violators in the most informative equilibrium when the opinion leader endorses the violation of the social norm (red dotted line), she does not endorse it (blue dashed line), and when she does not exist (solid gray line). The impact of the opinion leader's endorsement (lack thereof) is represented by the gap between the red dotted line and the gray solid line (the blue dashed line and the gray solid line). These

[11] The proof of Proposition 3 shows that the incentive compatibility constraint used to pin down $K^{\dagger}$ is the same as the one used to pin down $\overline{K}$.

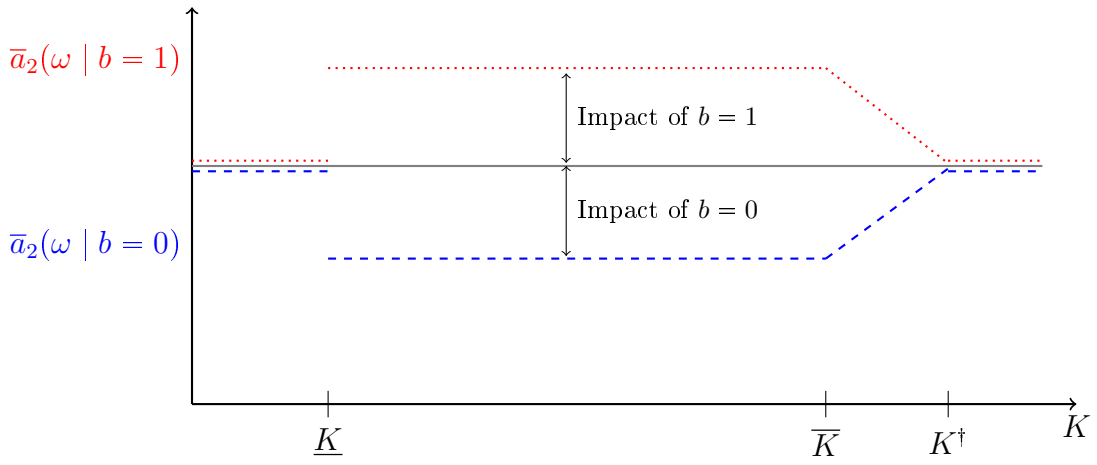

Figure 3: The impact of the opinion leader's endorsement decision.

Notes: The impact of the opinion leader on the share of violators in the most informative equilibrium when $\omega = 0$. The red dotted line shows the share of violators when the opinion leader endorses the violation ($b = 1$). The blue dashed line shows the share of violators when the opinion leader does not endorse the violation ($b = 0$). The gray solid line represents the share of violators when the opinion leader does not exist.

gaps exist only in the regions where informative equilibria exist. In a fully informative equilibrium the impact of the opinion leader is independent of her ideological strength, while in a partially informative equilibrium it decreases with her ideological strength.

Proposition 6. *The impact of the opinion leader's endorsement is increasing in α, λ and ψ, and it is decreasing in γ. The impact of the opinion leader's lack of endorsement exhibits the opposite comparative statics.*

When α and ψ increase, the opinion leader conveys information about a more relevant and uncertain variable. Her impact is then higher. When γ increases, the share of individuals with extreme preferences grows larger. These individuals do not respond to social concerns and the impact of the opinion leader thus decreases. Finally, when the norm entrenchment λ is high, a large mass of individuals in the traditional group abides by the norm due to social costs. In this case, the opinion leader's endorsement has a bigger potential to change societal behavior: her impact is higher.

5 Extensions

5.1 Homophily

In the baseline model, individuals match with each others with uniform probability. In several social interactions, however, there exists some degree of homophily: individuals with a preference toward (against) a social norm attend specific social environments and thus interact more often with individuals who share their preferences.

To capture this feature, assume that individuals with a positive (negative) private payoff θ_i are more likely to meet other individuals with positive (negative) private payoffs.[12] In particular, the probability with which an individual with a private payoff $\theta_i \in [-\gamma, \gamma]$ meets an individual with private payoff θ_j is given by:

$$m\left(\theta_j \mid \theta_i\right) = \begin{cases} \frac{1+h}{(2+h)\gamma} & \text{if } \theta_i\theta_j \geqslant 0; \\ \frac{1}{(2+h)\gamma} & \text{if } \theta_i\theta_j < 0. \end{cases}$$

The parameter $h \in \mathbb{R}_+$ measures the degree of homophily. When $h = 0$, there is no homophily and the model collapses to the baseline one. As h grows, the degree of homophily grows as well. In the limit as $h \to \infty$ individuals with positive (negative) private payoffs interact only with individuals with positive (negative) private payoffs.

Holding constant the expected shift in societal preferences, $\mathbb{E}[\omega \mid \mathcal{I}_2]$, an increase in the degree of homophily always increases the share of violators.[13] As h grows larger, individuals interact more often with like-minded individuals. The restraining power of the norm entrenchment is thus weaker and the share of violators goes up.

A straightforward adaptation of the proofs of Proposition 2 and Proposition 3 shows that a fully informative and a partially informative equilibrium exist as long as the ideological strength of the opinion leader is neither too high, nor too low. Furthermore, an increase in the degree of homophily shifts the range $[\underline{K}, \overline{K}]$ where a fully informative equilibrium exists to the left. When the degree of homophily increases, the share of violators increases too. This lowers the expected cost of the endorsement. Then, both

[12] We thus model type-dependent homophily. In our model, the behavior of other individuals is uncertain. Thus action-dependent homophily would be less reasonable (for a discussion of action vs. type-dependent homophily, see Bilancini et al., 2018). Finally, although we model homophily with two groups (those with private payoff greater than zero, and those with private payoff lower than zero), our arguments immediately generalize to settings in which the population is partitioned in any discrete number of groups representing connected intervals of the space of private payoffs.

[13] The comparative statics with respect to the other parameters are as in the baseline model.

$\underline{K}$ and $\overline{K}$ decrease. Finally, because individuals now interact more often with like-minded individuals, homophily amplifies the impact of the endorsement decision. Appendix B provides a formal statement and a proof of the results discussed in this section.

To sum up, as the society becomes more segregated in echo-chambers and individuals interact more often with other individuals sharing their preferences, the impact of opinion leaders grows larger. Moreover, the opinion leaders with the largest impact on society are those who are less ideological and more concerned about their popularity. This theoretical prediction is broadly in line with recent political events, namely the rise in ideological polarization paired with the success of political leaders characterized by wavering ideology and populist tendencies (e.g., Donald Trump in the US or Boris Johnson in the UK).

5.2 Uncertainty about the Opinion Leader's Type

In our baseline model, individuals know the ideological strength of the opinion leader; that is, they know $K \equiv k/(1-k)$. Often, however, individuals may be uncertain about the preferences of the opinion leader. The insights of our paper extend to this case. Suppose that individuals believe that the opinion leader's ideological strength is distributed in the interval $[K_\ell, K_h] \subset [0, +\infty)$ according to a continuously differentiable cdf G. Let g be the associated pdf that we assume strictly positive everywhere.[14]

The behavior of the opinion leader can thus be represented by a function $\beta : \{0,1\} \times [K_\ell, K_h] \to [0,1]$, where $\beta(s, K)$ is the probability of an endorsement when the opinion leader has ideological strength K and received signal s. As in the baseline model, the opinion leader is (weakly) more likely to endorse the violation of the norm after signal $s = 1$ than after signal $s = 0$. The opinion leader is also more likely to endorse the violation of the norm if her ideological strength is higher. To sum up, β is increasing in both its arguments.

We can summarize the optimal behavior of the opinion leader with a pair of thresholds (K_0^*, K_1^*): after signal s, the opinion leader endorses the violation of the norm if and only if her ideological strength is greater or equal to K_s^*. The previous discussion implies that $K_0^* \geqslant K_1^*$: if an opinion leader with ideological strength K endorses the violation of the norm after signal $s = 0$, she also endorses it after signal $s = 1$.

[14] Equivalently, we can assume that k in equation 3 is distributed in the interval $[k_\ell, k_h] \subset [0,1]$ according to a continuously differentialbe cdf.

In this setting an equilibrium exists. This equilibrium is informative as long as the ideological strength of the opinion leader is neither too high, nor too low. Finally, when the equilibrium is informative, the impact of the opinion leader is increasing in the probability that her ideological strength lies in the interval $[K_0^*, K_1^*]$.

The opinion leader has no impact when her ideological strength is excessively high.[15] In this case, she endorses the violation of the norm independently of the signal she received. The endorsement is thus ideologically motivated and lacks credibility. This case arises if $K_\ell > K_0^*$ even when the impact of the endorsement is null. As Appendix C shows, this happens if and only if:

$$K_\ell \geqslant \frac{1}{2\gamma - \lambda}\left(\gamma + \frac{2\gamma - \lambda}{2\gamma - \alpha\lambda} \cdot \frac{\alpha\psi}{3}\right) = K^\dagger \tag{10}$$

Inequality (10) defines the same threshold $K^\dagger$ we identified in the baseline model.[16]

The opinion leader has no impact also when her ideological strength is too low.[17] In this case, she does not endorse the violation of the norm due to the expected popularity cost that an endorsement causes. This case arises if $K_h < K_1^*$ even when the share of violators after the endorsement is maximal.[18] As Appendix C shows, this happens if and only if:

$$K_h < \frac{1}{2\gamma - \lambda}\left(\gamma - \frac{\alpha\psi}{3}\right) = \underline{K} \tag{11}$$

Inequality (11) defines the same threshold we identified in the baseline model.

The findings of our baseline model thus generalize to the case in which the ideological strength of the opinion leader is uncertain. When the range of possible ideological strengths, $[K_\ell, K_h]$, is either shifted to much to the right ($K_\ell > K^\dagger$) or too much to the left ($K_h < \underline{K}$), the opinion leader cannot convey any information and equilibria are uninformative. Instead, when the range $[K_\ell, K_h]$ includes intermediate values of K, the

[15] Formally, this requires the cdf on K to be equal to 0 at the two thresholds $G(K_0^*) = G(K_1^*) = 0$

[16] The uncertainty about K implies that all opinion leaders (but possibly a mass of measure zero) play a pure strategy. Unlike in the partially informative equilibrium of our baseline model, the uncertainty concerning the informational content of the endorsement decision comes from opinion leaders with different ideological strengths playing different pure actions, rather than from one opinion leader mixing.

[17] This requires the cdf on K to be equal to 1 at the two thresholds $G(K_0^*) = G(K_1^*) = 1$

[18] If the opinion leader never endorses the violation of the norm, beliefs after an endorsement are not pinned down. The expected share of violators after an endorsement, $\mathbb{E}[\overline{a}_2(\omega) \mid b = 1]$ is thus not pinned down either. However, the incentives to endorse the violation increase with $\mathbb{E}[\overline{a}_2(\omega) \mid b = 1]$. Hence, if there exists an equilibrium with no endorsements when $\mathbb{E}[\overline{a}_2(\omega) \mid b = 1]$ is maximal, there also exists an equilibrium with no endorsement when $\mathbb{E}[\overline{a}_2(\omega) \mid b = 1]$ is less than maximal.

opinion leader conveys some information. Proposition C.1 in Appendix C characterizes the impact of the opinion leader.

5.3 Multiple Opinion Leaders

Individuals often gather information from multiple sources: they may read several newspapers, listen to multiple pundits on TV, or follow different political leaders on social media. To capture this multiplicity, suppose individuals observe the endorsement decisions of two opinion leaders: $m \in \{1, 2\}$. Let $K_m = k_m/(1 - k_m)$ denote the ideological strength of opinion leader m.

Each opinion leader privately receives an independent signal about the societal change. The signal generating technology is the same for the two opinion leaders. Each opinion leader then decides independently and simultaneously whether to endorse the violation of the norm or not. The share of violators in the second period is still equal to the expression defined in Proposition 1, but the information set $\mathcal{I}_2^T$ now includes the endorsement decisions of both opinion leaders.

When multiple opinion leaders exist, the bounds for information transmission characterized in Proposition 2 and Proposition 3 still apply to each opinion leader. To see why, note that opinion leaders move independently and their payoffs (conditional on endorsing the violation) are linear in the share of violators. The law of iterated expectations thus implies that the expected payoff of each opinion leader is identical to the one in our baseline model. In particular, if $K_m \in [\underline{K}, \overline{K}]$ a fully informative equilibrium exists, while if $K_m \in (\overline{K}, K^\dagger)$ a partially informative equilibrium exists.

Although the existence of multiple opinion leaders does not affect the ability of each of them to convey information, it affects the impact they have on society. Indeed, the impact of an opinion leader now depends on what other opinion leaders do. For example, suppose the two opinion leaders play a fully informative strategy. If both opinion leaders endorse the violation or if they both do not, then their joint impact is larger (in either direction) than the one in the baseline model with only one opinion leader. Instead, when one opinion leader endorses the violation and the other does not, the overall impact is null and the share of violators is the one in Remark 1.[19]

Appendix D provides a more thorough analysis of the case with two opinion leaders. The same logic generalizes to the case in which more than two opinion leaders exist.

[19] Compared to Remark 1, individuals' beliefs are now less uncertain; that is, the variance of the posterior belief is lower. Due to risk neutrality, this lower variance is irrelevant.

6 Conclusions

Prominent political figures, popular media stars and successful social media influencers often affect the behavior of individuals through their actions, statements and endorsements. These opinion leaders modify societal behavior for better or for worse.

In this paper, we study when and to what extent an opinion leader can ease or hinder societal change by endorsing the violation of an established social norm. We build a model in which individuals with heterogeneous propensities to violate the norm are randomly matched and suffer a social cost if they choose to break the norm, while their match does not. A random shock modifies the propensity to abide by the norm among a group of individuals in the society. The majority of the society does not know the extent of this shock and may keep abiding by the norm due to social costs. An opinion leader who opposes the norm is partially informed about the shock. Her endorsement of the norm-violating behavior can inform individuals about the extent of the shock and turn some abiders into violators.

We show that the opinion leader's endorsement (or lack thereof) can shape societal behavior when she is neither too ideologically sided against the current norm, nor too popularity concerned. We also show that the impact of the opinion leader is larger in societies where the shock to societal preferences is more uncertain and affects a larger share of the population, and in societies where the entrenchment of the norm is higher. Furthermore, the impact of the opinion leader is higher in societies where individuals are more likely to interact with individuals who share a similar propensity to violate (or abide by) the social norm.

Our work highlights how the strategic incentives of opinion leaders affect their ability to ease or hinder societal change. It thus contributes to the current debate on the role and scope of prominent figures in shaping societies.

Appendix

A Proofs

Proof of Proposition 1

The cutoff strategies in period $t = 2$ are identified by a threshold in the novel group, $\theta_2^N(\omega)$, and a threshold in the traditional group, θ_2^T. Individuals above these threshold

are violators, while individuals below them are abiders. The threshold in the novel group thus solves:

$$\overline{\theta}_2^N(\omega) = \left(1 - (1-\alpha)\int_{\overline{\theta}_2^T}^{\gamma} \frac{dx}{2\gamma} - \alpha \int_{\overline{\theta}_2^N(\omega)}^{\omega+\gamma} \frac{dx}{2\gamma}\right)\lambda.$$

Solving and rearranging, we get:

$$\overline{\theta}_2^N(\omega) = \frac{\lambda}{2\gamma - \alpha\lambda}\left(\gamma + (1-\alpha)\overline{\theta}_2^T - \alpha\omega\right). \tag{A-1}$$

The share of violators in the novel group is thus state-dependent and equal to:

$$\overline{a}_2^N(\omega) = \int_{\overline{\theta}_2^N(\omega)}^{\omega+\gamma} \frac{dx}{2\gamma} = \frac{1}{2} - \frac{\lambda}{2\gamma(2\gamma - \alpha\lambda)}\left(\gamma + (1-\alpha)\overline{\theta}_2^T\right) + \frac{\omega}{2\gamma - \alpha\lambda}.$$

Now consider the threshold in the traditional group; it satisfies the following equation:

$$\overline{\theta}_2^T = \left(1 - (1-\alpha)\int_{\overline{\theta}_2^T}^{\gamma} \frac{dx}{2\gamma} - \alpha \int_{-\psi}^{\psi}\left(\int_{\overline{\theta}_2^N(\omega)}^{\omega+\gamma} \frac{dx}{2\gamma}\right) f(\omega \mid \mathcal{I}_2^T)d\omega\right)\lambda, \tag{A-2}$$

where $f(\omega \mid \mathcal{I}_2^T)$ is the pdf representing the posterior beliefs about ω by the individuals in the traditional group when their information set is $\mathcal{I}_2^T$. Note that:

$$\begin{aligned}\int_{-\psi}^{\psi}\left(\int_{\overline{\theta}_2^N(\omega)}^{\omega+\gamma} \frac{dx}{2\gamma}\right) f(\omega \mid \mathcal{I}_2^T)d\omega = \int_{-\psi}^{\psi} \overline{a}_2(\omega) f(\omega \mid \mathcal{I}_2^T)d\omega = \mathbb{E}\left[\overline{a}_2^N(\omega) \mid \mathcal{I}_2^T\right] = \\ = \frac{1}{2} - \frac{\lambda}{2\gamma(2\gamma - \alpha\lambda)}\left(\gamma + (1-\alpha)\overline{\theta}_2^T\right) + \frac{\mathbb{E}[\omega \mid \mathcal{I}_2^T]}{2\gamma - \alpha\lambda},\end{aligned}$$

where the last inequality follows from replacing for $\overline{a}_2^N(\omega)$. If we substitute this expression into equation (A-2) and rearrange, we get:

$$\overline{\theta}_2^T = \frac{\lambda\gamma}{2\gamma - \lambda} - \frac{\alpha\lambda}{2\gamma - \lambda}\mathbb{E}\left[\omega \mid \mathcal{I}_2^T\right]. \tag{A-3}$$

Plugging equation (A-3) into equation (A-1), we get the cutoff in the novel group:

$$\overline{\theta}_2^N(\omega) = \frac{\lambda\gamma}{2\gamma - \lambda} - \frac{\alpha\lambda}{2\gamma - \alpha\lambda}\left(\omega + \frac{\lambda(1-\alpha)}{2\gamma - \lambda}\mathbb{E}\left[\omega \mid \mathcal{I}_2^T\right]\right). \tag{A-4}$$

The shares of violators in the two groups are thus equal to:

$$\begin{aligned}
\overline{a}_2^T &= \int_{\overline{\theta}_2^T}^{\gamma} \frac{dx}{2\gamma} = \overline{a}_1 + \frac{\alpha\lambda}{2\gamma(2\gamma-\lambda)}\mathbb{E}\left[\omega \mid \mathcal{I}_2^T\right], \\
\overline{a}_2^N(\omega) &= \int_{\overline{\theta}_2^N(\omega)}^{\omega+\gamma} \frac{dx}{2\gamma} = \overline{a}_1 + \frac{1}{2\gamma-\alpha\lambda}\left(\omega + \frac{\lambda^2\alpha(1-\alpha)}{2\gamma(2\gamma-\lambda)}\mathbb{E}\left[\omega \mid \mathcal{I}_2^T\right]\right).
\end{aligned}$$

The overall share of violators follows from taking the weighted sum of these two expression with weights α and $1-\alpha$. □

Proof of Remark 1

When the opinion leader does not exist, individuals in the traditional group receive no information concerning ω. Hence, $\mathbb{E}\left[\omega \mid \mathcal{I}_2^T\right] = 0$. The shares of violators in the two groups become $\overline{a}_2^T = \overline{a}_1$ and $\overline{a}_2^N(\omega) = \overline{a}_1 + \frac{1}{2\gamma-\alpha\lambda}\omega$. The expression for $a_2^{NL}(\omega)$ follows from taking the weighted sum of these two quantities. □

Proof of Proposition 2

When the endorsement strategy is $(\beta(0), \beta(1)) = (0, 1)$, Bayes rule implies that the individuals' posterior beliefs about ω are equal to $f(\omega \mid b = 0) = \frac{1}{2\psi} - \frac{\omega}{2\psi^2}$ and $f(\omega \mid b = 1) = \frac{1}{2\psi} + \frac{\omega}{2\psi^2}$. This implies that the expected values of ω are equal to $\mathbb{E}[\omega \mid b = 0] = -\frac{\psi}{3}$ and $\mathbb{E}[\omega \mid b = 1] = \frac{\psi}{3}$. The share of violators in the society is thus given by equations (5) and (6) in the main text.

In a fully informative equilibrium, the opinion leader must endorse the violation of the norm after signal $s = 1$ (first inequality below), and refrain from doing so after signal $s = 0$ (second inequality):

$$\begin{aligned}
k - (1-k)\left(1 - \mathbb{E}\left[\overline{a}_1 + \frac{\alpha}{2\gamma-\alpha\lambda}\left(\omega + \frac{(1-\alpha)\lambda}{2\gamma-\lambda}\cdot\frac{\psi}{3}\right) \mid s = 1\right]\right) &\geqslant 0 \\
0 \geqslant k - (1-k)\left(1 - \mathbb{E}\left[\overline{a}_1 + \frac{\alpha}{2\gamma-\alpha\lambda}\left(\omega + \frac{(1-\alpha)\lambda}{2\gamma-\lambda}\cdot\frac{\psi}{3}\right) \mid s = 0\right]\right)
\end{aligned}$$

Substituting for $\mathbb{E}[\omega \mid s=0] = -\frac{\psi}{3}$ and $\mathbb{E}[\omega \mid s=1] = \frac{\psi}{3}$, and recalling that $K = \dfrac{k}{1-k}$, we can rewrite the two credibility constraints as:

$$K \geqslant \underline{K} = \frac{1}{2\gamma - \lambda}\left(\gamma - \frac{\alpha\psi}{3}\right) \tag{A-5}$$

$$K \leqslant \overline{K} = \frac{1}{2\gamma - \lambda}\left(\gamma + \frac{2(\gamma-\lambda)+\alpha\lambda}{2\gamma - \alpha\lambda} \cdot \frac{\alpha\psi}{3}\right) \tag{A-6}$$

These two inequalities define the range $[\underline{K}, \overline{K}]$ in the statement of the proposition. Given that $\gamma > \psi$, $\underline{K}$ is bounded above zero and $\underline{K} < \overline{K}$. Furthermore, $\overline{K}$ is also bounded above.

Finally, suppose that $K \in [\underline{K}, \overline{K}]$. It is immediate to see that $(\beta(0), \beta(1)) = (0,1)$ is optimal given the response of individuals specified by equations (5) and (6). Moreover, the cutoff strategies specified in the proof of Proposition 1 are optimal for all individuals when $(\beta(0), \beta(1)) = (0,1)$. □

Proof of Proposition 3

Suppose there exists an equilibrium in which the opinion leader adopts a partially informative strategy $(\beta(0), \beta(1)) \neq (0,1)$ with $\beta(0) < \beta(1)$. The expected value of ω conditional on the endorsement decision b would be:

$$\mathbb{E}[\omega \mid b=0] = -\frac{\beta(1)-\beta(0)}{2-\beta(1)-\beta(0)} \cdot \frac{\psi}{3} \qquad \text{and} \qquad \mathbb{E}[\omega \mid b=1] = \frac{\beta(1)-\beta(0)}{\beta(1)+\beta(0)} \cdot \frac{\psi}{3}.$$

In equilibrium, the opinion leader must be willing to choose $b=1$ with probability $\beta(1)$ after signal $s=1$, and to choose $b=1$ with probability $\beta(0)$ after signal $s=0$. Consider the case in which $\beta(0) \in (0,1)$ and $\beta(1) = 1$. The opinion leader must endorse the violation after $s=1$ and be indifferent between endorsing or not after signal $s=0$. The following two conditions must hold:

$$K \geqslant \frac{\gamma}{2\gamma-\lambda} - \frac{\alpha}{2\gamma-\alpha\lambda}\left(\frac{(1-\alpha)\lambda}{2\gamma-\lambda} \cdot \frac{1-\beta(0)}{1+\beta(0)} + 1\right) \cdot \frac{\psi}{3}$$

$$K = \frac{\gamma}{2\gamma-\lambda} - \frac{\alpha}{2\gamma-\alpha\lambda}\left(\frac{(1-\alpha)\lambda}{2\gamma-\lambda} \cdot \frac{1-\beta(0)}{1+\beta(0)} - 1\right) \cdot \frac{\psi}{3}.$$

From the equality, we get:

$$\beta(0) = 1 - 2 \cdot \frac{\frac{1}{\lambda(1-\alpha)}\left(1 - \frac{3(2\gamma-\alpha\lambda)}{\alpha\psi}\left(K - \frac{\gamma}{2\gamma-\lambda}\right)\right)(2\gamma-\lambda)}{1 + \frac{1}{\lambda(1-\alpha)}\left(1 - \frac{3(2\gamma-\alpha\lambda)}{\alpha\psi}\left(K - \frac{\gamma}{2\gamma-\lambda}\right)\right)(2\gamma-\lambda)}. \tag{A-7}$$

The right-hand side of equation (A-7) is increasing in K. Furthermore, $\beta(0) > 0$ if

$$K > \overline{K} = \frac{1}{2\gamma-\lambda}\left(\gamma + \frac{2(\gamma-\lambda)+\alpha\lambda}{2\gamma-\alpha\lambda} \cdot \frac{\alpha\psi}{3}\right)$$

and $\beta(0) < 1$ if

$$K < K^{\dagger} = \frac{1}{2\gamma-\lambda}\left(\gamma + \frac{2\gamma-\lambda}{2\gamma-\alpha\lambda} \cdot \frac{\alpha\psi}{3}\right).$$

This proves the existence of a partially informative equilibrium in which $\beta(0) \in (0,1)$ and $\beta(1) = 1$ for any $K \in (\overline{K}, K^{\dagger})$. If we substitute the $\beta(0)$ and $\beta(1)$ we just obtained in the overall share of violators, we obtain equations (7) and (8).

Now, suppose there exists an equilibrium in which $\beta(0) = 0$ and $\beta(1) \in (0,1)$. In this case, we would need:

$$K = \frac{\gamma}{2\gamma-\lambda} - \frac{\alpha}{2\gamma-\alpha\lambda}\left(\frac{(1-\alpha)\lambda}{2\gamma-\lambda} + 1\right) \cdot \frac{\psi}{3}$$
$$K \leqslant \frac{\gamma}{2\gamma-\lambda} - \frac{\alpha}{2\gamma-\alpha\lambda}\left(\frac{(1-\alpha)\lambda}{2\gamma-\lambda} - 1\right) \cdot \frac{\psi}{3}$$

Hence, a partially informative equilibrium exists if and only if K is non-generic and equal to $\underline{K} = \frac{1}{2\gamma-\lambda}\left(\gamma - \frac{\alpha\psi}{3}\right)$. □

Proof of Proposition 4

In an uninformative equilibrium, the expectations of individuals do not react to the endorsement decision of the opinion leader: $\mathbb{E}[\omega \mid b = 0] = \mathbb{E}[\omega \mid b = 1] = \mathbb{E}[\omega] = 0$. This happen when, in equilibrium, the opinion leader does not modify her behavior based on the signal she receives. We can thus have two possible scenarios.

First, the opinion leader may endorse the violation of the norm no matter which signal she received. Optimality requires $\beta(1) \geqslant \beta(0)$. This first scenario arises when the

opinion leader endorses the violation of the norm after signal $s = 0$; namely when:

$$K \geqslant \frac{1}{2\gamma - \lambda}\left(\gamma + \frac{2\gamma - \lambda}{2\gamma - \alpha\lambda}\frac{\alpha\psi}{3}\right) = K^{\dagger}.$$

Second, the opinion leader may not endorse the violation of the norm no matter which signal she received. Because optimality requires $\beta(0) < \beta(1)$, this second scenario arises when the opinion leader does not endorse the violation after signal $s = 1$, namely when

$$K \leqslant \frac{1}{2\gamma - \lambda}\left[\gamma - \frac{2\gamma - \lambda}{2\gamma - \alpha\lambda}\frac{\alpha\psi}{3}\right] := K^{U} \in (\underline{K}, \overline{K}).$$

Proof of Proposition 5

First, consider $\underline{K} = \frac{1}{2\gamma-\lambda}\left(\gamma - \frac{\alpha\psi}{3}\right)$. It is immediate to verify that this bound is decreasing in α and ψ, while it is increasing in λ. Given that

$$\frac{\partial \underline{K}}{\partial \gamma} = \frac{2\alpha\psi - 3\lambda}{3(2\gamma - \lambda)^2},$$

the bound is increasing in γ if $\alpha \geqslant \frac{3\lambda}{2\psi}$ and decreasing if the reversed inequality holds.

Now consider $\overline{K} = \frac{1}{2\gamma-\lambda}\left(\gamma + \frac{2(\gamma-\lambda)+\alpha\lambda}{2\gamma-\alpha\lambda} \cdot \frac{\alpha\psi}{3}\right)$. This bound is increasing in ψ. Consider the derivative of $\overline{K}$ with respect to α:

$$\frac{\partial \overline{K}}{\partial \lambda} = \frac{4\gamma^2(3\gamma - 3\alpha\lambda - \psi\alpha(1-\alpha)) + 4\alpha^2\psi\gamma(\gamma - \alpha) + \alpha^2\lambda^2(3\gamma + \psi(2-\alpha))}{3(2\gamma - \lambda)^2(2\gamma - \alpha\lambda)^2}.$$

Note that $4\gamma^2\left(3\gamma - 3\alpha\lambda - \psi\alpha(1-\alpha)\right) > 12\gamma^2(\gamma - \lambda/2 - \psi/4) > 0$, where the first inequality follows from $\alpha < 1/2$ and the second one from Assumption 1. We conclude that $\overline{K}$ is increasing in λ. The derivative with respect to α is equal to:

$$\frac{\partial \overline{K}}{\partial \alpha} = \frac{4\gamma^2 - \alpha^2\lambda^2 - 4(1-\alpha)\lambda\gamma}{3\alpha^2\lambda^2(2\gamma - \lambda) + 12\gamma(2\gamma - \lambda)(\gamma - \alpha\lambda)}\psi.$$

This expression is positive since both the numerator and the denominator are positive (see Assumption 1). Thus $\overline{K}$ increases with α. Finally consider the derivative with respect to γ:

$$\frac{\partial \overline{K}}{\partial \gamma} = -\frac{3\alpha^2\lambda^3 + 2\psi\alpha\lambda^2(2-\alpha^2) - 8\psi\alpha\gamma(1-\alpha)\lambda + 8\psi\alpha\gamma(\gamma - \lambda) + 12\lambda\gamma(\gamma - \alpha\lambda)}{3(2\gamma - \lambda)^2(2\gamma - \alpha\lambda)^2}.$$

As $12\lambda\gamma(\gamma-\alpha\lambda)-8\psi\alpha\gamma(1-\alpha)\lambda > 12\gamma\lambda\left(\gamma-\alpha\max\{\lambda,\psi\}\right) > 0$, this derivative is negative. The results on the range of values of K for which the fully informative equilibrium exists, $[\underline{K},\overline{K}]$, follow immediately from the previous analysis.

Now consider the partially informative equilibria. It is easy to verify that $K^\dagger = \frac{1}{2\gamma-\lambda}\left(\gamma+\frac{2\gamma-\lambda}{2\gamma-\alpha\lambda}\cdot\frac{\alpha\psi}{3}\right)$ is increasing in ψ and α. $K^\dagger$ is also increasing in λ and decreasing in γ. Indeed, by Assumption 1 we have

$$\frac{\partial K^\dagger}{\partial\lambda} = \frac{\alpha^2\gamma^2(3\lambda+\psi)+4\alpha^2\gamma\psi(\gamma-\lambda)+12\gamma^2(\gamma-\alpha\lambda)}{3(2\gamma-\lambda)^2(2\gamma-\alpha\lambda)^2} > 0,$$

while

$$\frac{\partial K^\dagger}{\partial\gamma} = -\frac{\alpha\gamma^2(3\alpha\lambda+2\psi)+8\alpha\gamma\psi(\gamma-\lambda)+12\gamma\lambda(\gamma-\alpha\lambda)}{3(2\gamma-\lambda)^2(2\gamma-\alpha\lambda)^2} < 0.$$

The results on the range of values of K for which a partially informative equilibrium exists, $[\overline{K},K^\dagger]$, follow from the derivatives computed above. □

Proof of Proposition 6

In a fully informative equilibrium, we have that:

$$\begin{aligned}\overline{a}_2(\omega \mid b=0)-\overline{a}_2^{NL}(\omega) &= -\frac{\alpha}{2\gamma-\alpha\lambda}\cdot\frac{(1-\alpha)\lambda\psi}{3(2\gamma-\lambda)}\\ \overline{a}_2(\omega \mid b=1)-\overline{a}_2^{NL}(\omega) &= \frac{\alpha}{2\gamma-\alpha\lambda}\cdot\frac{(1-\alpha)\lambda\psi}{3(2\gamma-\lambda)}\end{aligned}$$

Hence, in a fully informative equilibrium, the impact of the opinion leader's endorsement, $b=1$, on the share of violators is increasing in ψ and λ, while it is decreasing in γ. The impact of a lack of endorsement is reversed. Finally, the derivative of the opinion leader's impact with respect to α after an endorsement is equal to $\frac{\lambda\psi(2\gamma-4\alpha\gamma+\alpha^2\lambda)}{3(2\gamma-\alpha\lambda)^2(2\gamma-\lambda)}$. This expression is always positive because $\alpha<\frac{1}{2}$. The impact of the lack of an endorsement is symmetric with opposite sign.

In a partially informative equilibrium, instead, the impact of the opinion leader is asymmetric depending on whether she endorses the violation or not. If the opinion leader endorses the violation, her impact is

$$1-K-\frac{\gamma-\lambda}{2\gamma-\lambda}+\frac{\alpha}{2\gamma-\alpha\lambda}\cdot\frac{\psi}{3}.$$

By Assumption 1, this expression is increasing in α, λ and ψ, while it is decreasing in K and γ. If the opinion leader does not endorse the violation, her impact is

$$K - \frac{\gamma}{2\gamma - \lambda} - \frac{\alpha}{2\gamma - \alpha\lambda} \cdot \frac{\psi}{3}.$$

Again by Assumption 1, this expression is decreasing in α, λ and ψ, while it is increasing in K and γ. □

B Homophily: Formal Results

In this section we provide a formal statement (and the related proof) of the results on homophily discussed in Section 5.1.

Proposition B.1. *When social interactions are characterized by a degree of homophily equal to $h \in \mathbb{R}_+$, the share of violators in the first and second period are equal to*

$$\bar{a}_1(h) = \frac{2+h}{2}\left(\frac{\gamma - \lambda}{(2+h)\gamma - (1+h)\lambda}\right)$$
$$\bar{a}_2(\omega, h) = \bar{a}_1(h) + \frac{2+h}{2} \cdot \frac{\alpha}{(2+h)\gamma - (1+h)\alpha\lambda} \cdot \left(\omega + \frac{(1-\alpha)(1+h)\lambda}{(2+h)\gamma - (1+h)\lambda}\mathbb{E}\left[\omega \mid \mathcal{I}_2\right]\right)$$

Both these shares are increasing in h and so it is the impact of the opinion leader. Furthermore, as h increases the range $[\underline{K}, \overline{K}]$ for which a fully informative equilibrium exists shifts to the left, while the range $(\overline{K}, K^\dagger)$ for which a partially informative equilibrium exists can either shift to the left or widen.

Proof. The same logic used in the proof of Proposition 1 implies that when the matching probabilities are given by $m(\cdot \mid \cdot)$ the share of violators in the first period is equal to

$$\overline{a}_1(h) = \frac{(2+h)}{2}\left(\frac{\gamma - \lambda}{(2+h)\gamma - (1+h)\lambda}\right).$$

The derivative of $\overline{a}_1(h)$ with respect to h is equal to: $\frac{\partial \overline{a}_1(h)}{\partial h} = \frac{\lambda(\gamma-\lambda)}{2((2+\gamma)h-(1+h)\lambda)^2}$, which is positive because $\gamma > \lambda$.

In the second period we can again define two cutoff strategies. The traditional group adopts a state-independent cutoff strategy with thresholds $\overline{\theta}_2^T(h)$. The novel group adopts a state-dependent cutoff strategy with threshold $\overline{\theta}_2^N(\omega, h)$. The two thresholds

are given by:

$$\begin{aligned}\overline{\theta}_2^T(h) &= \frac{\lambda\gamma}{(2+h)\gamma-(1+h)\lambda} - \frac{(1+h)\alpha\lambda}{(2+h)\gamma-(1+h)\lambda}\mathbb{E}\left[\omega \mid \mathcal{I}_2^T\right] \\ \overline{\theta}_2^N(\omega,h) &= \frac{\lambda\gamma}{(2+h)\gamma-(1+h)\lambda} - \\ &\quad - \frac{(1+h)\alpha\lambda}{(2+h)\gamma-(1+h)\alpha\lambda}\left(\omega + \frac{(1-\alpha)(1+h)\lambda}{(2+h)\gamma-(1+h)\lambda}\mathbb{E}\left[\omega \mid \mathcal{I}_2^T\right]\right)\end{aligned}$$

The share of violators in the two groups are then given by:

$$\begin{aligned}\overline{a}_2^T(h) &= \int_{\bar{\theta}_2^O(h)}^{\gamma}\frac{dx}{2\gamma} = \bar{a}_1 + \frac{1}{2\gamma}\cdot\frac{\lambda(1+h)\alpha}{(2+h)\gamma-(1+h)\lambda}\mathbb{E}\left[\omega \mid \mathcal{I}_2^T\right] \\ \overline{a}_2^N(\omega,h) &= \int_{\bar{\theta}_2^N(\omega,h)}^{\omega+\gamma}\frac{dx}{2\gamma} = \overline{a}_1(h) \\ &\quad + \frac{1}{(2+h)\gamma-(1+h)\alpha\lambda}\left(\frac{2+h}{2}\omega + \frac{\alpha(1-\alpha)(1+h)^2\lambda^2}{2\gamma\left[(2+h)\gamma-(1+h)\lambda\right]}\mathbb{E}\left[\omega \mid \mathcal{I}_2^T\right]\right)\end{aligned}$$

The overall share of violators is obtained taking the weighted sum of $\overline{a}_2^T(h)$ and $\overline{a}_2^N(\omega,h)$:

$$\begin{aligned}\overline{a}_2(\omega,h) &= (1-\alpha)\overline{a}_2^T(h) + \alpha\overline{a}_2^N(\omega,h) = \overline{a}_1(h) \\ &\quad + \frac{2+h}{2}\cdot\frac{\alpha}{(2+h)\gamma-(1+h)\alpha\lambda}\cdot\left(\omega + \frac{(1-\alpha)(1+h)\lambda}{(2+h)\gamma-(1+h)\lambda}\mathbb{E}\left[\omega \mid \mathcal{I}_2^T\right]\right)\end{aligned}$$

This share is increasing in the degree of homophily. To see why, recall that $\overline{a}_1(h)$ is increasing in h. Then observe that the derivatives with respect to h of the terms in ω and in $\mathbb{E}\left[\omega \mid \mathcal{I}_2^T\right]$ are proportional to ω and $\mathbb{E}\left[\omega \mid \mathcal{I}_2^T\right]$. Hence, the overall derivative with respect to h is minimized when $\omega = \mathbb{E}\left[\omega \mid \mathcal{I}_2^T\right] = -\psi$. This minimal value is positive. Hence, $\overline{a}_2(\omega,h)$ is increasing in h. The same argument also proves that the impact of the opinion leader is increasing in h.

If we replicate the steps of the proof of Proposition 2 and we take into account that in a fully informative equilibrium we still have $\mathbb{E}[\omega \mid b=0] = -\frac{\psi}{3}$ and $\mathbb{E}[\omega \mid b=1] = \frac{\psi}{3}$, one obtains that a fully informative equilibrium exists if and only if $K \in [\underline{K}(h), \overline{K}(h)]$,

where:

$$\underline{K}(h) = \frac{1}{2\left(2+h\right)\gamma - \left(1+h\right)\lambda]}\left(\left(2+h\right)\gamma - h\lambda - \frac{\left(2+h\right)\alpha\psi}{3}\right)$$

$$\overline{K}(h) = \frac{1}{2\left[\left(2+h\right)\gamma - \left(1+h\right)\lambda\right]}\Bigg(\left(2+h\right)\gamma - h\lambda + \\ + \frac{\left(2+h\right)\gamma - \left(1+h\right)\lambda - \left(1+h\right)\left(1-\alpha\right)\lambda}{\left(2+h\right)\gamma - \left(1+h\right)\alpha\lambda} \cdot \frac{\left(2+h\right)\alpha\psi}{3}\Bigg).$$

The derivative of $\underline{K}(h)$ with respect to h is given by:

$$\frac{\partial \underline{K}(h)}{\partial h} = -\lambda \frac{3\left(\gamma - \lambda\right) + \alpha\psi}{6\left(\left(2+h\right)\gamma - \left(1+h\right)\lambda\right)^2} < 0$$

while the derivative of $\overline{K}(h)$ with respect to h is given by:

$$\frac{\partial \overline{K}(h)}{\partial h} = -\lambda \frac{3(\gamma-\lambda) - \alpha\psi}{6[(2+h)\gamma - (1+h)\lambda]} - \frac{\alpha\lambda\psi}{3} \frac{(1-\alpha)[\gamma^2(4(1+h)+h^2) - \alpha\lambda^2(1+h)^2]}{[(2+h)\gamma - (1+h)\lambda]^2[(2+h)\gamma - (1+h)\alpha\lambda]^2}.$$

This expression is bounded above by

$$-\frac{\alpha\lambda\psi}{6[(2+h)\gamma - (1+h)\lambda]^2} \frac{2(2+h)\gamma - (1+h)(1+\alpha)}{[(2+h)\gamma - (1+h)\alpha\lambda]^2} \alpha\lambda^2(1+h),$$

which is negative. Hence, both $\underline{K}(h)$ and $\overline{K}(h)$ are decreasing in h.

Now consider partially informative equilibria in which $\beta(1) = 1$ and $\beta(0) \in (0,1)$. The same steps of Proposition 3 imply that the upper bound $K^{\dagger}(h)$ is equal to:

$$K^{\dagger}(h) = \frac{1}{2}\left[1 + \frac{\lambda}{(2+h)\gamma - (1+h)\lambda} + \frac{(2+h)\alpha\psi}{3[(2+h)\gamma - (1+h)\alpha\lambda]}\right].$$

The derivative of this expression with respect to h can be either negative or positive depending on whether $(\gamma-\lambda)/[(2+h)\gamma - (1+h)\lambda]^2$ is greater or lower than $\alpha\psi/[(2+h)\gamma - (1+h)\alpha\lambda]^2$. □

C Uncertainty about the Opinion Leader's Ideological Strength: Formal Results

In this section we provide a formal statement (and the related proof) of the results about the robustness of our insights to the case in which the ideological strength of the opinion leader is uncertain (see Section 5.2).

Proposition C.1. *Suppose that the ideological strength of the opinion leader is distributed in the interval* $[K_\ell, K_h] \subset [0, +\infty)$ *according to a continuously differentiable cdf* G*. Then, an equilibrium exists and it is characterized by the pair* (K_0^*, K_1^*)*. The expected share of violators in the second period is*

$$\overline{a}_2(\omega \mid b = 0) = \overline{a}_1 + \frac{\alpha}{2\gamma - \alpha\lambda}\left(\omega - \frac{(1-\alpha)\lambda}{2\gamma - \lambda} \cdot \frac{G(K_0^*) - G(K_1^*)}{2 - G(K_0^*) - G(K_1^*)} \cdot \frac{\psi}{3}\right)$$
$$\overline{a}_2(\omega \mid b = 1) = \overline{a}_1 + \frac{\alpha}{2\gamma - \alpha\lambda}\left(\omega + \frac{(1-\alpha)\lambda}{2\gamma - \lambda} \cdot \frac{G(K_0^*) - G(K_1^*)}{2 - G(K_0^*) - G(K_1^*)} \cdot \frac{\psi}{3}\right)$$

The impact of the opinion leader is increasing in $G(K_0^*) - G(K_1^*)$*, the probability mass with which the ideological strength of the opinion leader is in-between* K_0^* *and* K_1^**.*

Proof. Suppose individuals believe the opinion leader is following the threshold strategies (K_0^*, K_1^*) defined in the main text and assume that a positive mass of opinion leaders do not endorse the violation after signal $s = 0$. By Bayes rule, we have:

$$\mathbb{E}[\omega \mid b = 0, K_0^*, K_1^*] = -\frac{G(K_0^*) - G(K_1^*))}{2 - G(K_0^*) - G(K_1^*)} \cdot \frac{\psi}{3} \tag{C-1}$$
$$\mathbb{E}[\omega \mid b = 1, K_0^*, K_1^*] = \frac{G(K_0^*) - G(K_1^*))}{2 - G(K_0^*) - G(K_1^*)} \cdot \frac{\psi}{3} \tag{C-2}$$

The expected share of violators from the opinion leader's point of view is thus equal to (see Proposition 1)

$$\mathbb{E}[\overline{a}_2(\omega) \mid s] = \begin{cases} \overline{a}_1 - \frac{\alpha}{2\gamma - \alpha\lambda}\left(1 + \frac{(1-\alpha)\lambda}{2\gamma - \lambda} \cdot \frac{G(K_0^*) - G(K_1^*)}{2 - G(K_0^*) - G(K_1^*)}\right)\frac{\psi}{3} & \text{if } s = 0 \\ \overline{a}_1 + \frac{\alpha}{2\gamma - \alpha\lambda}\left(1 + \frac{(1-\alpha)\lambda}{2\gamma - \lambda} \cdot \frac{G(K_0^*) - G(K_1^*)}{2 - G(K_0^*) - G(K_1^*)}\right)\frac{\psi}{3} & \text{if } s = 1 \end{cases} \tag{C-3}$$

Opinion leaders with ideological strengths equal to the cutoffs (K_0^*, K_1^*) must be indifferent between endorsing the violation of the norm and not doing so. The thresholds

thus jointly solve:

$$K_0^* = 1 - \overline{a}_1 + \frac{\alpha}{2\gamma - \alpha\lambda}\left(1 + \frac{(1-\alpha)\lambda}{2\gamma - \lambda} \cdot \frac{G(K_0^*) - G(K_1^*)}{2 - G(K_0^*) - G(K_1^*)}\right)\frac{\psi}{3} \tag{C-4}$$

$$K_1^* = 1 - \overline{a}_1 - \frac{\alpha}{2\gamma - \alpha\lambda}\left(1 + \frac{(1-\alpha)\lambda}{2\gamma - \lambda} \cdot \frac{G(K_0^*) - G(K_1^*)}{2 - G(K_0^*) - G(K_1^*)}\right)\frac{\psi}{3} \tag{C-5}$$

This is a system of two equations in two unknowns. Instead of writing this system in terms of ideological strengths, we could write it in terms of the original payoffs. To this goal, let H be the cdf of k in $[0,1]$, h be the associated pdf, and k_0^* and k_1^* be the relevant thresholds in terms of private payoffs. The system then becomes:

$$k_0^* = \frac{1 - \overline{a}_1 + \frac{\alpha}{2\gamma - \alpha\lambda}\left(1 + \frac{(1-\alpha)\lambda}{2\gamma - \lambda} \cdot \frac{H(k_0^*) - H(k_1^*)}{2 - H(k_0^*) - H(k_1^*)}\right)\frac{\psi}{3}}{2 - \overline{a}_1 + \frac{\alpha}{2\gamma - \alpha\lambda}\left(1 + \frac{(1-\alpha)\lambda}{2\gamma - \lambda} \cdot \frac{H(k_0^*) - H(k_1^*)}{2 - H(k_0^*) - H(k_1^*)}\right)\frac{\psi}{3}}$$

$$k_1^* = \frac{1 - \overline{a}_1 - \frac{\alpha}{2\gamma - \alpha\lambda}\left(1 + \frac{(1-\alpha)\lambda}{2\gamma - \lambda} \cdot \frac{H(k_0^*) - H(k_1^*)}{2 - H(k_0^*) - H(k_1^*)}\right)\frac{\psi}{3}}{2 - \overline{a}_1 - \frac{\alpha}{2\gamma - \alpha\lambda}\left(1 + \frac{(1-\alpha)\lambda}{2\gamma - \lambda} \cdot \frac{H(k_0^*) - H(k_1^*)}{2 - H(k_0^*) - H(k_1^*)}\right)\frac{\psi}{3}}$$

The right-hand side of the previous system is a continuous function that maps a convex and compact space, $[0,1] \times [0,1]$, into itself. By the Brouwer fixed point theorem, the system (hence, the original one defined in terms of K) has an equilibrium. Every pair (K_0^*, K_1^*) that satisfies the system is a solution. It is also immediate to verify that $K_0^* \geqslant K_1^*$.

Equation (C-3) implies that the impact of the opinion leader is larger when the probability mass of opinion leaders with ideological strength in-between K_0^* and K_1^* is larger. The bound (10) in the main text follows from (C-4) after we set $G(K_0^*) = G(K_0^*) = 0$. The bound (11), instead, follows from (C-5) assuming that the share of violators after the opinion leader's endorsement is maximal (i.e., it is equal to $1 + \frac{\alpha}{2\gamma - \alpha\lambda}\frac{(1-\alpha)\lambda}{2\gamma - \lambda}\frac{\psi}{3}$). □

D Multiple Opinion Leaders

Assume that two opinion leaders exist. The share of violators in the second period is still defined by the expression defined in Proposition 1. However, the information available to individuals now includes the endorsement behavior of both opinion leaders.

Since opinion leaders move independently, each opinion leader can play an uninformative strategy, a fully informative strategy, or a partially informative strategy.

Consider opinion leader 1 (the analysis for opinion leader 2 is identical and omitted). Conditional on the signal s that she received, her updated beliefs about the state ω are still as in the baseline model:

$$f(\omega \mid s=0) = \frac{1}{2\psi} - \frac{\omega}{2\psi^2} \qquad \text{and} \qquad f(\omega \mid s=1) = \frac{1}{2\psi} + \frac{\omega}{2\psi^2}.$$

Hence, if she received signal $s=0$, her expectation about ω is $-\psi/3$. Moreover, the probability she assigns to opinion leader 2 having received signal $s=1$ is given by:

$$\int_{-\psi}^{\psi} \left(\frac{1}{2\psi} - \frac{\omega}{2\psi^2}\right)\left(\frac{1}{2} + \frac{\omega}{2\psi}\right) d\omega = \frac{1}{3}.$$

Instead, if she received signal $s=1$, the expected value of ω is $\psi/3$ and the probability she assigns to opinion leader 2 also receiving signal $s=1$ is:

$$\int_{-\psi}^{\psi} \left(\frac{1}{2\psi} + \frac{\omega}{2\psi^2}\right)\left(\frac{1}{2} + \frac{\omega}{2\psi}\right) d\omega = \frac{2}{3}.$$

Clearly, if opinion leader 2 is playing an uninformative equilibrium strategy, opinion leader 1 is in a situation that is analogous to the one characterized in the main text. Thus, the results in Propositions 2 and 3 still apply.

Suppose that opinion leader 2 is playing a fully informative strategy, $(\beta(0), \beta(1)) = (0,1)$. If opinion leader 1 also plays a fully informative strategy, individuals can face one of 4 possible pairs of endorsement decisions. The expectations of individuals would react to each possibile pair of endorsements as summarized by the following table:

1 \ 2	$b_2 = 0$	$b_2 = 1$
$b_1 = 0$	$-\frac{\psi}{2}$	0
$b_1 = 1$	0	$\frac{\psi}{2}$

Table D1: $\mathbb{E}[\omega \mid \cdot]$ given opinion leaders' endorsement decisions.

Hence, if both opinion leaders play a fully informative strategy, the expected expectation of the individuals from the point of view of the opinion leader is $-\frac{\psi}{2} \cdot \frac{2}{3} = -\frac{\psi}{3}$ after signal $s=0$ and $\frac{\psi}{2} \cdot \frac{2}{3} = \frac{\psi}{3}$ after signal $s=1$.

The expected payoff of opinion leader 1 when she she has received signal $s = 1$, she believes opinion leader 2 is playing a fully informative strategy, and she chooses $b = 1$ is equal to:

$$k_1 + (1 - k_1)\left[1 - a_2^{FI}(\omega \mid b = 1)\right]$$

where $a_2^{FI}(\omega \mid b = 1)$ is defined in equation (6). Proceeding as in the proof of Proposition 2 we conclude that truthful information transmission after $s = 1$ is incentive compatible for the opinion leader if and only if $K_1 > \underline{K}$. A similar reasoning also implies that truthful information transmission is incentive compatible for the opinion leader when she receives signal $s = 0$ if and only if $K_1 < \overline{K}$.

The same logic implies that if $K_1 \in (\overline{K}, K^{\dagger})$, opinion leader 1 can play a partially informative equilibrium (see the proof of Proposition 3 for details).

Finally, suppose that opinion leader 2 plays a partially informative strategy in which $\beta(0) \in (0, 1)$ and $\beta(1) = 1$. As in the previous case, the law of iterated expectation implies that the expected expectation of individuals from the point of view of opinion leader 1 is equal to the one in the baseline model. Hence, the expected payoff of opinion leader 1 remains unchanged and the bounds we derived in Propositions 2 and 3 still apply.

The share of violators is obtained by replacing the relevant expectations (e.g., the expressions in Table D1) in the expression for $\overline{a}_2(\omega)$ in Proposition 1.

References

Acemoglu, D. and M. O. Jackson (2015). History, expectations, and leadership in the evolution of social norms. *The Review of Economic Studies 82*(2), 423–456.

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

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

Alonso, R. and O. Câmara (2016). Persuading voters. *American Economic Review 106*(11), 3590–3605.

Baron, D. P. (2006). Persistent media bias. *Journal of Public Economics 90*(1), 1–36.

Benabou, R. and J. Tirole (2011). Laws and norms. *NBER working paper No. 17579*.

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

Bilancini, E., L. Boncinelli, and J. Wu (2018). The interplay of cultural intolerance and action-assortativity for the emergence of cooperation and homophily. *European Economic Review 102*, 1–18.

Bursztyn, L., G. Egorov, and S. Fiorin (2020). From extreme to mainstream: The erosion of social norms. *American Economic Review 110*(11), 3522–48.

Carlsson, M., G. B. Dahl, and D.-O. Rooth (2016). Do politicians change public attitudes? *IZA Discussion Paper No. 10349*.

Chan, J., S. Gupta, F. Li, and Y. Wang (2019). Pivotal persuasion. *Journal of Economic theory 180*, 178–202.

Che, Y.-K., W. Dessein, and N. Kartik (2013). Pandering to persuade. *American Economic Review 103*(1), 47–79.

Cowen, T. and D. Sutter (1998). Why only nixon could go to china. *Public Choice 97*(4), 605–615.

Cukierman, A. and M. Tommasi (1998). When does it take a nixon to go to china? *The American Economic Review 88*(1), 180–197.

Fainmesser, I. P. and A. Galeotti (2021). The market for online influence. *American Economic Journal: Microeconomics 13*(4), 332–372.

Friedrichsen, J., T. König, and T. Lausen (2021). Social status concerns and the political economy of publicly provided private goods. *The Economic Journal 131*(633), 220–246.

Gallice, A. and E. Grillo (2020). Economic and social-class voting in a model of redistribution with social concerns. *Journal of the European Economic Association 18*(6), 3140–3172.

Gentzkow, M. and J. M. Shapiro (2006). Media bias and reputation. *Journal of Political Economy 114*(2), 280–316.

Gerardi, D., E. Grillo, and I. Monzón (2022). The perils of friendly oversight. *Journal of Economic Theory*, 105500.

Gratton, G. (2014). Pandering and electoral competition. *Games and Economic Behavior 84*, 163–179.

Grosjean, P. A., F. Masera, and H. Yousaf (2021). Whistle the racist dogs: Political campaigns and police stops.

Gulotty, R. and Z. Luo (2021). Fire alarm fatigue: How politicians evade accountability. *Available at SSRN 3815391*.

Hinnosaar, M. and T. Hinnosaar (2021). Influencer cartels. *Available at SSRN*.

Hopkins, E. and T. Kornienko (2004). Running to keep in the same place: Consumer choice as a game of status. *American Economic Review 94*(4), 1085–1107.

Jackson, M. O. and X. Tan (2013). Deliberation, disclosure of information, and voting. *Journal of Economic Theory 148*(1), 2–30.

Levy, G. and R. Razin (2015). Preferences over equality in the presence of costly income sorting. *American Economic Journal: Microeconomics 7*(2), 308–37.

Loeper, A., J. Steiner, and C. Stewart (2014). Influential opinion leaders. *The Economic Journal 124*(581), 1147–1167.

Maskin, E. and J. Tirole (2019). Pandering and pork-barrel politics. *Journal of Public Economics 176*, 79–93.

Mitchell, M. (2021). Free ad (vice): Internet influencers and disclosure regulation. *The RAND Journal of Economics 52*(1), 3–21.

Morelli, M. and R. Van Weelden (2013). Ideology and information in policymaking. *Journal of Theoretical Politics 25*(3), 412–439.

Morris, S. (2001). Political correctness. *Journal of Political Economy 109*(2), 231–265.

Mullainathan, S. and A. Shleifer (2005). The market for news. *American Economic Review 95*(4), 1031–1053.

Müller, K. and C. Schwarz (2020). From hashtag to hate crime: Twitter and anti-minority sentiment. *Available at SSRN 3149103*.

Ottaviani, M. and P. N. Sørensen (2006a). Professional advice. *Journal of Economic Theory 126*(1), 120–142.

Ottaviani, M. and P. N. Sørensen (2006b). Reputational cheap talk. *The Rand Journal of Economics 37*(1), 155–175.

Prat, A. and D. Strömberg (2013). The political economy of mass media. *Advances in Economics and Econometrics 2*, 135.

Prato, C. and I. R. Turner (2022). Institutional foundations of the power to persuade. *Center for Open Science SocArXiv No. 4w9af*.

Schnakenberg, K. E. (2015). Expert advice to a voting body. *Journal of Economic Theory 160*, 102–113.

Schnakenberg, K. E. (2017). Informational lobbying and legislative voting. *American Journal of Political Science 61*(1), 129–145.