The Polarization Paradox: Why More Connections Can Divide Us*

Arthur Campbell† C. Matthew Leister‡ Philip Ushchev§ Yves Zenou¶

June 3, 2025

Abstract

We develop a simple model to examine how social network structure shapes political polarization. Despite its tractability, the model generates a rich set of equilibrium behaviors. First, we show that moderate views, expressed by centrist contents, benefit from a pairwise advantage and is amplified when it is the most common type. However, when extreme views dominate, sufficiently dense networks can amplify polarization. Second, we establish a non-monotonic relationship between network density and polarization, offering a theoretical explanation for the mixed empirical evidence on the role of social media. Third, we show that partial polarization is less likely in nearly regular networks than in those with more dispersed degree distributions. We also explore two extensions: one where agents can abstain from forwarding extreme content, and another where ideological preferences are asymmetric. In both cases, the results remain, but the conditions for full or no polarization become more restrictive.

JEL classification numbers: D83, D85, L82.

Key words: Social networks, network density, content filtering, polarization.

*We thank the participants of the Asian-Pacific Industrial Organisation Society Annual Conference (Melbourne, December 2018), the Fifth Annual Network Science and Economics Conference (Bloomington, April 2019), the 24th Coalition Theory Network Workshop (Aix-en-Provence, May 2019), the 2021 Cambridge-INET Networks Webinar Series, the departmental seminars at the University of Queensland, the University of Hong Kong, and the University of Antwerp, for their helpful comments. The financial support from the Australian Research Council (DP200102547) is gratefully acknowledged.

†Monash University, Australia. E-mail: arthur.campbell@monash.edu.

‡Monash University, Australia. E-mail: matthew.leister@monash.edu.

§ECARES, ULB, Belgium. E-mail: ph.ushchev@gmail.com.

¶Monash University, Australia, and CEPR. E-mail: yves.zenou@monash.edu.

1 Introduction

The recent rise of populism and political polarization has generated widespread concern about how political opinions form and evolve, as well as their implications for democratic institutions. In response, a growing literature has examined the role of social networks in shaping political beliefs and behaviors. A prevalent view links the rise of polarization to the increasing use of social media and denser patterns of online interaction, which are said to facilitate the formation of "echo chambers"—environments in which individuals are primarily exposed to ideologically aligned content (see, e.g., Mutz, 2006; Hindman, 2009; Pariser, 2011; Sustein, 2008; Sunstein, 2018). These dynamics are thought to reinforce prior beliefs and limit exposure to opposing views, thereby intensifying polarization (e.g., El-Bermawy, 2016; Allcott and Gentzkow, 2017; Allcott et al., 2020).

However, a more recent wave of empirical studies challenges this narrative, arguing that social media can, under certain conditions, mitigate polarization. By expanding users' exposure beyond immediate social circles to include weak ties and diverse viewpoints, platforms like Twitter and Facebook may counteract ideological narrowing (Barberá, 2015; Barberá et al., 2015; Boxell et al., 2017; Barberá, 2020; Dubois and Blank, 2018; Algan et al., 2025). The coexistence of these two perspectives—with empirical evidence supporting both—highlights the need for a more nuanced theoretical understanding of how network structure influences polarization.

This paper contributes to this debate by developing a theoretical framework that links the density and heterogeneity of social networks to the degree of ideological polarization. Specifically, we ask: Does greater network connectivity (i.e., denser social communication) always lead to greater polarization, or can it instead foster moderation? We address this question by building a dynamic model of content diffusion on random networks. Our approach draws on the theory of random graphs (Newman et al., 2001, 2002) and builds on recent applications in economics (Campbell, 2013; Campbell et al., 2024).

To the best of our knowledge, our model is the first to systematically explore the interplay between network structure and polarization in a dynamic setting with endogenous behavior. Despite its simplicity, the model yields rich and realistic outcomes. In particular, we uncover a *non-monotonic* relationship between network density and polarization: intermediate levels of connectivity amplify polarization, while very sparse or very dense networks tend to promote centrist behavior. This theoretical result helps reconcile the conflicting findings in the empirical literature.

Our framework features a population of agents aligned with one of three ideological types—left, centrist, and right—positioned along a one-dimensional spectrum. Inspired by classical models of spatial competition in political economy and industrial organization, we assume that individuals prefer content matching their own ideology but rank centrist content higher than the opposite extreme. Content diffuses through peer-to-peer recommendations, and the composition of forwarded content evolves endogenously over time.

We find that when centrist agents are sufficiently prevalent, the equilibrium amplifies moderate views due to a pairwise advantage: centrist content is never more than one ideological step away from any agent. However, when extremists dominate the population, intermediate levels of network density can reinforce their influence and lead to full polarization. Strikingly, when the network becomes very dense, exposure again becomes representative of the overall population, restoring the prevalence of moderate content.

We derive several novel results. First, we identify the thresholds at which polarization, partial polarization, or full polarization becomes the dominant behavioral outcome. Second, we show that the relationship between network density and polarization is highly nonlinear: polarization is most acute in networks that are neither too sparse nor too dense. Third, we analyze the effects of a mean-preserving spread in the degree distribution and find that polarization is more likely in networks with greater variance in connectivity. This aligns with empirical concerns about "super-spreaders" or influencers disproportionately shaping ideological content.

To assess the robustness of our results, we consider two key extensions. First, we allow agents to abstain from forwarding content when exposed only to ideologically distant messages. Even with this behavioral refinement, we find that polarization emerges under intermediate connectivity and dispersed preference distributions. Second, we introduce asymmetry in the ideological composition of the population and show that small deviations from symmetry shrink the regions of full or zero polarization, leading to a wider domain of partial polarization.

Overall, our model provides a tractable yet flexible framework for understanding how network structure shapes ideological outcomes. It delivers new theoretical insights into when communication fosters consensus and when it fuels polarization, offering a unified explanation for the mixed empirical evidence on the effects of social media on political discourse.

2 Related literature

This paper contributes to the growing literature on echo chambers, political polarization, and the role of network structure in shaping collective behavior. We build on and connect several strands of work—both theoretical and empirical—that explore how individual exposure to content, preferences, and network connectivity interact to determine ideological outcomes.

2.1 Related empirical literature

A large body of empirical research examines the informational consequences of algorithmic curation and social network structure. Ross Arguedas et al. (2022) provide an extensive review of studies on echo chambers, filter bubbles, and selective exposure, concluding that interventions aimed at diversifying content exposure—such as reducing algorithmic homophily—do not consistently reduce polarization. Rather, outcomes depend on the structure of exposure and the cognitive and emotional responses of users. Similarly, Terren and Borge (2021) emphasize that content amplification dynamics, rather than mere exposure patterns, are crucial to understanding polarization. These findings support a key feature of our model: agents do not passively absorb what they observe; they actively choose what to forward based on their ideological preferences.

Experimental evidence also points to unintended consequences of increased exposure to opposing views. Bail et al. (2018) conduct a randomized experiment in which Twitter users followed bots that introduced them to opposing ideological content. Instead of reducing polarization, the intervention increased it—especially among Republicans—highlighting the potential for backfire effects. Levy (2021) similarly find that moderate increases in exposure to diverse views on social media can increase polarization due to backlash or motivated reasoning, although at very high exposure levels, moderation may occur.

Network structure also plays a pivotal role. Barberá et al. (2015) show that users embedded in denser political discussion networks tend to be more ideologically extreme, though some cross-cutting ties can moderate attitudes—albeit under specific conditions. Finally, Guess et al. (2023) use a large-scale field experiment during the 2020 U.S. election to test the effects of altering feed algorithms on Facebook and Instagram. They find that reducing exposure to like-minded content had no significant impact on polarization or political attitudes, suggesting that connectivity or structural changes alone may not yield predictable effects.

Our contribution. Our model formalizes and unifies these empirical insights by introducing a dynamic framework in which agents repeatedly forward content based on their ideological preferences and social exposure. We characterize the steady-state distribution of forwarded content and show that it depends nonlinearly on both the degree distribution and the ideological composition of the population.

First, we identify threshold effects in the relationship between network density and behavioral polarization. For a fixed centrist share, there exists a critical level of network density below which centrist content dominates and above which extremist content can be amplified. Second, when the centrist population is small, intermediate levels of connectivity can generate maximal polarization—even when centrist agents are numerous. Third, as network density becomes very large, the behavioral distribution converges to the population's ideological composition, as agents begin to sample the full diversity of preferences.

Our model is, to our knowledge, the first to demonstrate that increasing network density can have *non-monotonic* effects on polarization. Sparse networks favor centrism, since extremists are less likely to encounter reinforcing content. Intermediate density enables ideological sorting and amplification, while high density reintroduces moderation as exposure becomes representative. These findings yield new, testable implications about when connectivity fosters or mitigates polarization and complement empirical work by offering microfounded behavioral mechanisms.

2.2 Related theoretical literature

Our paper also contributes to the theoretical literature on political polarization and opinion dynamics in networked settings.

Levy and Razin (2019) review foundational mechanisms behind echo chambers, emphasizing how cognitive frictions—such as correlation neglect—can produce polarization even in the absence of extreme preferences. Their work motivates our dynamic approach, which emphasizes how behavioral responses to social exposure evolve over time. Callander and Carbajal (2022) develop a dynamic model where polarization emerges endogenously through feedback loops between media content and voter behavior. While their focus is on strategic media-voter interactions, our model centers on peer-to-peer content diffusion and its dependence on network structure.

Della Lena et al. (2023) examine affective polarization in a repeated setting where individuals interact through media and social channels. Our model differs in focusing on ideological (rather than emotional) alignment and on how content preferences shape the flow of information. Finally, Bolletta and Pin (2025) study dynamic opinion formation with endogenous network evolution. While they show how homophily and clustering can endogenously produce polarization, we instead hold the network exogenous and vary its degree distribution parametrically. This allows us to isolate how the structural properties of the network alone affect the equilibrium distribution of ideological behavior.

Our contribution. We introduce a dynamic model of ideological content diffusion with exogenous but variable network structure. Agents forward content aligned with their type, and behavior evolves based on what content is observed and passed on. We show that polarization responds *non-monotonically* to network density: Centrism prevails in sparse and dense networks, but intermediate connectivity amplifies extremism. This result, absent in prior models, offers a new explanation for when network expansion can unintentionally fuel polarization—despite increased exposure.

The rest of the paper is organized as follows. Section 3 introduces our baseline model and characterizes the steady-state equilibria. Section 4 presents the comparative statics with respect to the degree distribution and establishes our main result: denser networks have a non-linear effect on polarization. In Section 5, we explore two extensions of the baseline model. In the first (Section 5.1), agents are allowed to abstain from forwarding content. In the second (Section 5.2), we introduce asymmetry in the distribution of extremists on the left and right. Section 6 concludes. All proofs are provided in the Appendix, with additional results available in the Online Appendix.

3 The baseline model

In this section, we describe and study our baseline model of content filtering and diffusion which we use to study the impact of network structure and preference distribution on polarization.

We characterize the conditions under which the social transmission mechanism amplifies centrist views, and those under which it favors the extremes. In the latter case, we refer to the outcome as *political polarization*—a state in which left- and right-leaning ideologies dominate public discourse at the expense of mainstream views.

3.1 Notation, definitions, and preferences

Consider a model of ideological competition along a one-dimensional discrete spectrum consisting of three content types: L ("left"), M ("middle"), and R ("right"). These correspond to political ideologies, with the centrist type M having a relative advantage in pairwise comparisons against the extremes. The population has unit mass, initially we assume that it is symmetrically composed of a measure ρ of individuals of type M, and a measure $(1-\rho)/2$ each of types L and R, with $0 < \rho < 1$.

We consider a dynamic model of content filtering with discrete time periods $t = 1, 2, \ldots$. In each period, all individuals select k friends from the previous period, where k is drawn from a degree distribution $\{p_k\}$ with expectation $\mu_1 = \mathbb{E}[k]$ satisfying $p_k > 0$ for some $k \geq 2$. In each period, an individual receives one content recommendation from each of their k friends from the previous period. They observe all recommended content and subsequently recommend the one that best aligns with their own ideological type to their friends in the next period. This dynamic process is what we term *content filtering*.

An individual's ranking of content types depends on their own ideological type, as summarized below:

Type	Ranking
L	$L \succ M \succ R$
M	$M \succ L \sim R$
R	$R \succ M \succ L$

Each agent strictly prefers content that matches their own type. In addition, a type-L or type-R individual prefers centrist (M) content to content from the opposing extreme. A type-M individual, by contrast, is indifferent between L and R content. An agent will forward content of their own type if they receive at least one such recommendation from their k friends. If they receive no content of their own type, they follow their preference ranking as given above to select among the recommended contents. In the case of a type-M individual receiving only L and R content (and thus being indifferent), they choose between the two with equal probability, i.e., 50–50.[1] We see that under these preferences type M content has a relative advantage in pairwise comparisons against either of the extremes.

[1] In the baseline model, we assume that agents are required to recommend some content, even if it does not align with their preferences. In Section 5.1, we relax this assumption and allow agents—particularly those who are ideologically distant from all received content (e.g., left-wing individuals receiving only right-wing content)—the option to remain silent and refrain from forwarding any political content. We also allow M-type individuals to remain silent if none of their friends made a recommendation in the previous period.

3.2 Content Filtering

We are interested in the steady-state distribution of political content recommendations across types as $t \to \infty$. We first develop a dynamic equation to describe the evolution recommended content. In each period t, every individual of any type selects k friends uniformly at random from the population in the previous period $t-1$. Let $y_t \in [0,1]$ denote the share of middle-oriented (centrist type M) individuals at time t. Accordingly, the population—normalized to have total mass one—comprises a fraction y_t of centrists, and equal fractions $\frac{1-y_t}{2}$ of left-wing and right-wing individuals, respectively.[2] This implies that the share of centrists y_t evolves according to

$$y_t = \sum_k p_k \left(\rho \left[1 - (1-y_{t-1})^k\right] + (1-\rho) \left[\left(\frac{1+y_{t-1}}{2}\right)^k - \left(\frac{1-y_{t-1}}{2}\right)^k \right] \right), \tag{1}$$

where $\{p_k\}$ is the degree distribution of the social network, with $k \in \mathbb{N}$ and ρ is the fraction of centrist-type individuals.

The first term, $\rho\left[1-(1-y_{t-1})^k\right]$, represents the probability that an individual is of type M and receives at least one recommendation of their most preferred content type (M) from k friends drawn uniformly at random from the population at time $t-1$. The second term, $(1-\rho)\left[\left(\frac{1+y_{t-1}}{2}\right)^k - \left(\frac{1-y_{t-1}}{2}\right)^k\right]$, is the probability that an individual is of type L or R and ends up recommending content of type M. This occurs when the individual hears neither: (i) content of their own (most preferred) type, nor (ii) exclusively content of their least preferred type (i.e., type R for a type-L individual, and type L for a type-R individual) from their k randomly selected friends.

It is convenient to rewrite this dynamic equation in terms of generating functions (Newman et al., 2001; Vega-Redondo, 2007; Newman, 2010; Campbell, 2013; Campbell et al., 2024). Specifically, equation (1) can be expressed as

$$y_t = f(y_{t-1}, \rho), \tag{2}$$

where the function $f : [0,1] \times [0,1] \to [0,1]$ is defined by

$$f(y,\rho) := \rho\left[1 - G(1-y)\right] + (1-\rho)\left[G\left(\frac{1+y}{2}\right) - G\left(\frac{1-y}{2}\right)\right], \tag{3}$$

and $G : [0,1] \to [0,1]$ is the generating function associated with the degree distribution $\{p_k\}$, given by

$$G(x) := \sum_{k \in \mathbb{N}} p_k x^k. \tag{4}$$

[2]This symmetry greatly simplifies the analysis. In Section 5.2, we relax this assumption and consider asymmetric cases in which the fractions of left- and right-wing individuals may differ.

3.3 Steady state equilibrium

We define $y^* \in [0,1]$ as a *steady state* of the dynamical system (2) if and only if it satisfies the fixed-point equation

$$y^* = f\left(y^*, \rho\right). \tag{5}$$

Equation (5) represents the steady-state condition derived from the dynamic equation (2) by omitting the time index t.

Define the following two threshold values:[3]

$$\rho_0 := \frac{1-\psi}{\mu_1 - \psi}, \tag{6}$$

$$\rho_1 := \frac{\mu_1 + p_1 - 2}{\mu_1 - p_1}, \tag{7}$$

where

$$\psi := \mathbb{E}\left[\widetilde{k}\left(\frac{1}{2}\right)^{\widetilde{k}-1}\right] = \sum_{k=1}^{\infty} k p_k \left(\frac{1}{2}\right)^{k-1} = G'\left(\frac{1}{2}\right), \tag{8}$$

is the *50%-discounted average degree.* The value ψ reflects the expected degree of a randomly chosen node, weighted by the probability that a neighbor relays a signal with probability $\frac{1}{2}$ at each hop. It captures how effectively information spreads in the network starting from a single source, discounting by distance.

The following proposition provides an exhaustive characterization of the stable steady states.

Proposition 1. *The steady-state equilibria defined by* (5) *are characterized as follows:*

*(i) If $\rho \leq \rho_0$, the unique stable steady state corresponds to **full polarization**, that is, $y^* = 0$.*

*(ii) If $\rho_0 < \rho < \rho_1$, the unique stable steady state corresponds to **partial polarization**, that is, $y^* \in (0,1)$. Moreover, y^* is strictly increasing in ρ, i.e., $\frac{\partial y^*}{\partial \rho} := y^*_\rho > 0$.*

*(iii) If $\rho \geq \rho_1$, the unique stable steady state corresponds to **no polarization**, that is, $y^* = 1$.*

This proposition identifies three distinct regimes for the long-run population state y^* based on the value of the key parameter ρ, which captures the relative strength of centrist influence.

When ρ is small (i.e., $\rho \leq \rho_0$), centrist influence is weak relative to the structure of the network, and the system converges to full polarization: all agents adopt extreme views ($y^* = 0$). In this regime, extremist content dominates due to two reinforcing factors: a scarcity of centrist types and the mechanics of diffusion in the network. From a behavioral standpoint, even centrist agents who only receive extreme content from their neighbors are more likely to propagate it, leading to an amplification of polarized behavior throughout the network.

In the intermediate regime ($\rho_0 < \rho < \rho_1$), the dynamics admit a unique interior steady state ($y^* \in (0,1)$), reflecting a polarized but heterogeneous society. This result implies that

[3] In Online Appendix A, we show that $\rho_0 \leq \rho_1$, with "$<$" if $p_k > 0$, for some $k \geq 3$.

the population contains enough centrist individuals to sustain some centrist behavior in the long run, but the network is not sufficiently supportive to eliminate polarization altogether. This reflects a society where different content types coexist in equilibrium: some individuals continue to endorse moderate views, while others promote extremes. The fact that $y_\rho^* > 0$ implies that even small increases in the prevalence of centrist preferences lead to a monotonic rise in centrist behavior—suggesting a smooth, continuous policy response.

Finally, when ρ is large ($\rho \geq \rho_1$), centrist influence dominates, and the society converges to consensus around the centrist position ($y^* = 1$). Under these conditions, centrist content fully displaces the extremes. This corresponds to an information environment in which the most socially acceptable behavior becomes centrist due to its overwhelming representation and central location in content space.

The thresholds ρ_0 and ρ_1 thus act as bifurcation points, demarcating qualitative changes in the equilibrium outcome. This characterization highlights how network structure and relative influence interact to shape long-run ideological composition. Thus, the threshold ρ_0 can be interpreted as a critical value for ρ below which the steady state $y^* = 0$ becomes stable. It delineates a regime where information diffusion fails to sustain itself due to insufficient influence. The threshold ρ_1 indicates a tipping point above which the steady state $y^* = 1$ becomes stable. Beyond this point, initial support for the behavior (or content) becomes self-reinforcing and widespread adoption (or consensus) can be sustained in the population.

These results illustrate that steady-state outcomes are highly nonlinear in ρ: small changes near the thresholds ρ_0 or ρ_1 can trigger qualitative shifts in long-run behavior. This has important implications for policy: interventions that raise centrist prevalence slightly above ρ_0 can avert full polarization, while pushing it above ρ_1 may achieve full consensus. The thresholds themselves depend on the network's degree distribution, emphasizing that structural properties of social networks shape collective outcomes.

Figure 1 illustrates how the distribution of preferences maps into the distribution of behavior,[4] where behavior reflects content recommendation patterns. As shown in the figure, the central position of the middle-type content in the product space leads to its amplification when preferences are either concentrated or uniformly distributed. In contrast, when the preference distribution is sufficiently dispersed—such that extreme preferences dominate—the extreme content types may be amplified at the expense of the middle type.

4 Comparative statics with respect to the degree distribution

Our broader goal is to understand how structural features of the social network affect the ideological landscape in the long-run. In particular, we aim to address the following questions:

- Does increasing network density lead to more or less polarization?

[4] Here, behavior refers to the act of recommending a particular type of content.

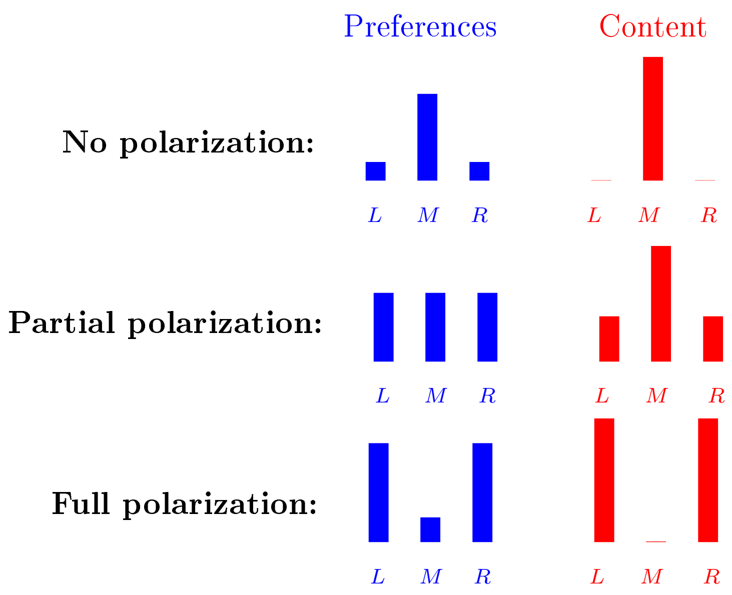


Figure 1: The mapping of the preference distribution into the behavior distribution

- Does heterogeneity in individuals' connectivity amplify or mitigate polarization?

We now turn to the comparative statics of the stable equilibrium with respect to the degree distribution of the network.

First, we analyze how increased network density affects polarization. By network density, we refer to changes in the degree distribution that are ordered according to the monotone likelihood ratio (MLR) criterion. Second, we derive sharper results by focusing on a parametric family of degree distributions—specifically, the exponential distribution—where density can be varied systematically. Third, we investigate how a mean-preserving spread in the degree distribution affects polarization, comparing the outcomes in more regular versus more dispersed (irregular) networks.

4.1 The impact of a denser network: General results

We define a *denser network* as one in which the degree distribution shifts in a way that the post-shock distribution dominates the pre-shock distribution according to the MLR ordering. Formally, consider a parametric family of degree distributions $\{p_k(\theta)\}_{k\in\mathbb{N}}$, indexed by a scalar parameter $\theta \in [0, \overline{\theta}]$, where $\overline{\theta}$ may be finite or infinite. We interpret higher values of θ as corresponding to denser networks.

We impose the following assumptions on the parametric family $\{p_k(\theta)\}$:

Assumption 1. *The family $\{p_k(\theta)\}$ is ordered by the monotone likelihood ratio (MLR). That is, for any k, the ratio $\frac{p_k(\theta')}{p_k(\theta)}$ is increasing in k whenever $\theta' > \theta$.*

Assumption 1 formalizes the idea of a *denser network* in terms of the monotone likelihood ratio (MLR) ordering. Economically, this means that higher values of θ correspond to network structures where individuals are more likely to have more connections, in a way that preserves a coherent shift in connectivity across the population. This ordering captures a form of network thickening that is behaviorally plausible. Thus, when Assumption 1 holds, we refer to θ as the *network density parameter*. Higher values of θ correspond to stochastically denser networks in the MLR sense.

The next two assumptions impose mild regularity conditions on the behavior of the distribution $\{p_k(\theta)\}$ as the density parameter θ approaches its extreme values: either a minimally connected network as $\theta \to 0$, or a maximally connected network as $\theta \to \overline{\theta}$.

Assumption 2. *As $\theta \to 0$, the degree distribution converges to that of a regular network in which each individual has exactly one connection:*

$$\lim_{\theta \to 0} p_k(\theta) = \begin{cases} 1, & k = 1; \\ 0, & k \geq 2. \end{cases} \tag{9}$$

As $\theta \to \overline{\theta}$, the network becomes infinitely dense in the sense that, for all $n \in \mathbb{N}$,

$$\lim_{\theta \to \overline{\theta}} \mathbb{P}[\widetilde{k} < n] = 0, \tag{10}$$

where $\widetilde{k}$ denotes the random degree of an individual drawn from the population.

Assumption 2 provides meaningful benchmarks at the two extremes of network density. As $\theta \to 0$, the network becomes minimally connected, converging to a *regular network* where every individual has only one connection. This represents an isolated or highly fragmented society with minimal social influence. As $\theta \to \overline{\theta}$, the network becomes infinitely dense—reflecting a society with extremely high connectivity.

Assumption 3. *As $\theta \to 0$, the degree distribution admits the following local approximation:*

$$p_1(\theta) = 1 - \lambda\theta + o(\theta), \tag{11}$$

$$p_2(\theta) = \lambda\theta + o(\theta), \tag{12}$$

$$\mathbb{P}[\widetilde{k} \geq 3] = o(\theta), \tag{13}$$

for some constant $\lambda > 0$.

Assumption 3 offers a second-order approximation of the degree distribution in sparse networks, capturing its local behavior as $\theta \to 0$. In this regime, the distribution is concentrated on degrees 1 and 2: most individuals have a single connection, while the probability of higher degrees vanishes more rapidly. The parameter λ governs the initial rate at which the network begins to densify. This assumption facilitates tractability of limit cases and helps characterize how small increases in connectivity near $\theta = 0$ influence aggregate behavior.

As the next result shows, the stable equilibrium y^* can exhibit pronounced monotonic responses to increases in network density.

Proposition 2. *Consider a parametric family of degree distributions $\{p_k(\theta)\}$ satisfying Assumptions 1-3. Then:*

(i) For each $\rho \in (0,1)$, there exists a threshold $\widehat{\theta}(\rho)$ such that

$$\theta < \widehat{\theta}(\rho) \quad \Longleftrightarrow \quad y^*(\theta) = 1 \quad \textit{(no polarization)}.$$

(ii) There is a threshold $\widehat{\rho} \in (0,1)$ such that, for all $\rho < \widehat{\rho}$, there exists a non-degenerate interval (θ_1, θ_2) with $0 < \theta_1 < \theta_2 < \overline{\theta}$ such that

$$\theta \in (\theta_1, \theta_2) \quad \Longrightarrow \quad y^*(\theta) = 0 \quad \textit{(full polarization)}.$$

(iii) As the network becomes infinitely dense, the equilibrium content distribution converges to the share of centrists in the population:

$$\lim_{\theta \to \overline{\theta}} y^*(\theta) = \rho.$$

Proposition 2 provides key insights into how the structure (i.e., density) of social networks affects long-run ideological behavior in a dynamic recommendation model. Agents repeatedly share content aligned with their ideological preferences, but their behavior is shaped by the recommendations they receive through the network. The degree distribution governs how many friends (and hence recommendations) each agent receives, making it a central determinant of content diffusion.

Part (*i*) indicates that when the network is sufficiently sparse (i.e., $\theta < \widehat{\theta}(\rho)$), the unique steady state is *no polarization* ($y^* = 1$): all agents ultimately forward centrist (M) content. This occurs because in sparse networks, agents are more likely to receive few recommendations. For centrist agents, it is relatively easy to encounter M content, which they prefer and will forward. For extremist agents (L or R), receiving their own type becomes less likely, and the fallback behavior involves forwarding M content (since L prefers M to R and vice versa). Thus, centrist content gradually dominates. This reflects the idea that *low levels of social interaction can stabilize moderate behavior* even when extremist preferences are present in the population.

Part (*ii*) reveals that, when centrists are sufficiently rare ($\rho < \widehat{\rho}$), there exists an intermediate density range (θ_1, θ_2) in which extremist content dominates in equilibrium ($y^* = 0$). As the network becomes denser, extremist agents are more likely to receive content of their own type and propagate it. Centrist agents—although present—may become overwhelmed by L and R recommendations and end up forwarding non-centrist content due to lack of type-M contents. The result is *behavioral divergence* from the underlying preference distribution: A minority of extremists can dominate observed behavior through network-driven amplification. This captures how *increased social exposure can lead to echo chambers and radicalization* when centrists are underrepresented.

Finally, part (*iii*) shows that as the network becomes infinitely dense (i.e., $\theta \to \overline{\theta}$), each agent receives a content distribution that approximates the population type distribution. Thus, equilibrium behavior distribution converges to preference distribution: the steady-state share of forwarded centrist content equals the share of centrist agents, $y^*(\theta) \to \rho$. This represents an environment with perfect information flow, where agents are no longer biased by local sampling or social bottlenecks. Behavior and ideology align in aggregate, and network structure ceases to distort diffusion.

These results show that increasing network connectivity can have *non-monotonic* and counterintuitive effects. In sparse or fully connected networks, moderate content prevails. But in intermediate regimes—particularly when centrists are scarce—networks may fuel polarization. This highlights the need for careful design of communication platforms and interventions that account for both preference heterogeneity and network topology.[5]

4.2 The impact of a denser network: Exponential distribution

The findings of Proposition 2 are robust in that they do not depend on a particular parametric form of the degree distribution $\{p_k(\theta)\}$. However, to obtain sharper and more transparent results, it is useful to consider specific parametric families. One such example is the exponential degree distribution, given by

$$p_k = (1-\theta)\theta^{k-1}, \quad k \in \mathbb{N}, \tag{14}$$

where $\theta \in (0,1)$ governs the network density.

This family satisfies Assumptions 1–3: As θ increases, the network becomes denser in the monotone likelihood ratio (MLR) sense, and the limiting behavior is well-defined at both endpoints. Thus, θ serves as a valid and analytically convenient network density parameter. The next proposition provides a complete characterization of how polarization responds to network density under exponential degree distributions.

Proposition 3. *Assume that the degree distribution is exponential. Then:*

(i) If $\rho < \frac{1}{9}$, there exist three thresholds $0 < \theta_1 < \theta_2 < \theta_3 < 1$ such that the stable equilibrium

[5] In the Online Appendix C, we show how the change of three different degree distributions p_k^0, p_k^1, and p_k^2, such that $\{p_k^0\} \prec_{\text{FOSD}} \{p_k^1\} \prec_{\text{FOSD}} \{p_k^2\}$, leads to partial polarization under $\{p_k^0\}$, no polarization under $\{p_k^1\}$, and partial polarization again under $\{p_k^2\}$. This is another illustration of polarization being highly non-monotone with respect to network density.

$y^*(\theta, \rho)$ *satisfies:*

$$\begin{aligned} \theta \le \theta_1 &\implies y^*(\theta) = 1 \quad \textit{(no polarization)}, \\ \theta_1 < \theta < \theta_2 &\implies 0 < y^*(\theta) < 1 \quad \textit{(partial polarization)}, \\ \theta_2 \le \theta \le \theta_3 &\implies y^*(\theta) = 0 \quad \textit{(full polarization)}, \\ \theta > \theta_3 &\implies 0 < y^*(\theta) < 1 \quad \textit{(partial polarization)}, \\ \theta \to 1 &\implies y^*(\theta) \to \rho. \end{aligned}$$

(ii) If $\frac{1}{9} \le \rho < \sqrt{5} - 2 \approx 0.236$*, then* $y^*(\theta) = 1$ *when* θ *is close to zero. As* θ *increases,* $y^*(\theta)$ *varies non-monotonically: it first decreases, then increases, but remains strictly positive throughout. Thus, full polarization never arises.*

(iii) If $\rho \ge \sqrt{5} - 2$*, then* $y^*(\theta) = 1$ *for all* θ *below a certain threshold, and decreases with* θ *thereafter, but full polarization* ($y^* = 0$) *is never reached.*

Proposition 3 provides a rich bifurcation structure of equilibrium behavior as a function of the network density parameter θ, fully endogenized through the exponential degree distribution. The result in part (i) is especially striking: for low centrist shares ($\rho < 1/9$), polarization evolves *non-monotonically* with θ through four distinct behavioral regimes. As θ increases, the system starts in a consensus state favoring centrists ($y^* = 1$), then enters a partial polarization regime, transitions into full polarization ($y^* = 0$), and finally reverts to partial polarization, before converging smoothly to the ideological composition of the population ($y^* \to \rho$). This "double non-monotonicity" illustrates the complexity of the dynamic system. For intermediate and high values of ρ, the equilibrium is better behaved: full polarization is avoided, and y^* remains positive or decreasing but bounded away from zero.

The proposition also reveals how the interaction between population ideology (ρ) and network density (θ) governs the diffusion of political content. When centrists are rare, small networks favor centrist behavior simply because extremist agents lack enough exposure to reinforcing content. As the network grows denser, extremist types benefit disproportionately from improved connectivity, leading to amplification of their own content—and potentially crowding out centrist messages entirely. Yet, at very high levels of density, exposure becomes representative of the population, and behavior begins to reflect the underlying ideological mix, restoring proportionality. For larger centrist shares, the system exhibits more stability, and polarization is never complete. This has important implications for policy: Attempts to increase connectivity or exposure must consider where the system is positioned relative to these thresholds.

At the micro level, the results reflect how agents behave when receiving ideologically mixed content. In sparse networks, centrist agents reliably forward M content, and extremist agents often fall back on M when their own type is absent. This drives consensus around centrism. In moderately dense networks, however, extremists are more likely to receive their own type and propagate it, while centrists may be overwhelmed by polarized content, diluting their impact. At high density, every agent sees a broad ideological sample and forwards content that reflects their own type, yielding aggregate behavior consistent with the population's ideological composition.

Overall, Proposition 3 demonstrates how a simple change in the structure of social interactions—via the degree distribution—can produce complex, discontinuous shifts in ideological behavior, even when preferences are fixed.

Figure 2 illustrates the core findings of Proposition 3, demonstrating how the equilibrium level of centrist content $y^*(\theta)$ varies with network density θ under exponential degree distribution, across different values of the centrist population share ρ. Panel 2a shows the case where $\rho = 0.1 < \widehat{\rho} = 1/9$. Here, the function $y^*(\theta)$ is *non-monotonic*, featuring four distinct regimes: full centrism at low density, partial polarization, full polarization over a nontrivial interval, and partial polarization again at high density, eventually converging to ρ as $\theta \to 1$. This illustrates the double non-monotonicity discussed in Proposition 3(i) where there is full polarization when the network is dense enough but not too dense.

Panel 2b depicts the critical threshold case $\rho = \widehat{\rho} = 1/9$. The equilibrium curve becomes tangent to the horizontal axis without reaching full polarization. That is, $y^*(\theta)$ dips close to zero but remains strictly positive, marking the boundary between regimes with and without full polarization. This aligns with Proposition 3(ii), where the equilibrium is non-monotonic but avoids full polarization. Panels 2c and 2d illustrate how increasing the centrist share ($\rho > \widehat{\rho}$) stabilizes the system. As ρ increases, the dip in $y^*(\theta)$ becomes shallower and the function becomes strictly decreasing (or weakly non-monotonic) without reaching zero. These cases confirm that for larger centrist populations, network density continues to influence polarization, but full polarization is no longer sustainable.

Together, the panels in Figure 2 underscore the non-monotonic relationship between connectivity and ideological behavior. They show that while sparse and very dense networks promote centrism, intermediate levels of density can temporarily amplify polarization—particularly when the population is ideologically imbalanced.

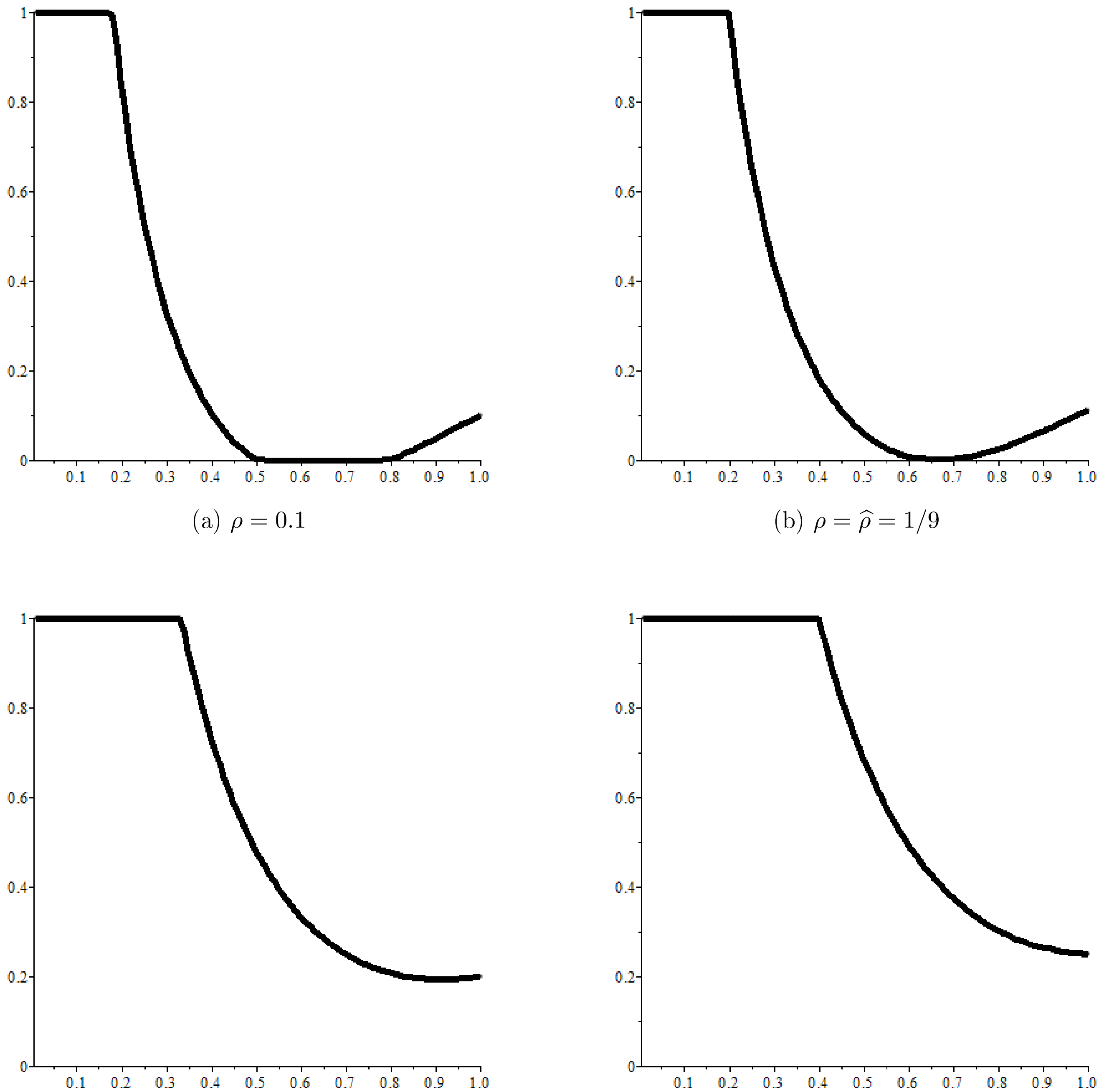
4.3 The impact of a mean-preserving spread on polarization

We now examine how the shape of the degree distribution—holding its mean fixed—affects the equilibrium polarization outcome. The following proposition focuses on a mean-preserving spread of the degree distribution.

Proposition 4. *Consider a mean-preserving spread of the degree distribution in which p_1 remains approximately constant. Then the polarization thresholds are such that ρ_0 decreases and ρ_1 increases.*

Recall that the thresholds ρ_0 and ρ_1 are given by (6) and (7), that is,

$$\rho_0 = \frac{1-\psi}{\mu_1 - \psi}, \quad \rho_1 = \frac{\mu_1 + p_1 - 2}{\mu_1 - p_1},$$



(a) $\rho = 0.1$

(b) $\rho = \widehat{\rho} = 1/9$

(c) $\rho = 0.2$

(d) $\rho = 0.25$

Figure 2: Equilibrium behavior $y^*(\theta)$ as a function of network density θ under exponential degree distribution

where $\mu_1 = \mathbb{E}[\widetilde{k}]$ is the mean degree and $\psi = \mathbb{E}[\widetilde{k} \cdot (1/2)^{\widetilde{k}-1}]$ is the discounted degree. A mean-preserving spread increases the variance of $\widetilde{k}$ while keeping μ_1 constant. This leads to a lower ψ, since larger values of k are more heavily discounted. As a result, ρ_0 decreases and ρ_1 increases, widening the range of ρ for which the equilibrium y^* lies strictly between 0 and 1.

Proposition 4 shows that increasing the dispersion of the degree distribution—without altering its mean—expands the intermediate region of the parameter space that supports *partial polarization*. That is, a broader degree distribution (e.g., one with more high-degree and low-degree individuals and fewer in the middle) increases the gap between ρ_0 and ρ_1. In contrast, more regular (less dispersed) networks make partial polarization less likely and push the system toward either full centrism ($y^* = 1$) or full polarization ($y^* = 0$). This is because a mean-preserving spread creates a more heterogeneous social environment: Some individuals become much more connected (potential influencers), while others remain marginally connected. This heterogeneity introduces asymmetry in the flow of content: Highly connected extremists can disproportionately amplify their views, while low-degree centrists have limited ability to counteract this influence. This result implies that structural inequality in social exposure—while keeping average exposure fixed—can increase the prevalence of partial polarization. This aligns with empirical concerns about "super-spreaders" in online networks and the asymmetric reach of ideological content. In contrast, more egalitarian networks (e.g., regular graphs) tend to generate corner outcomes, where the system either fully polarizes or fully moderates depending on the underlying ideological distribution.

5 Extensions

In this section, we present two extensions of our baseline model to demonstrate the robustness of our main results. In Subsection 5.1, we allow agents to refrain from recommending their least preferred content. In Subsection 5.2, we consider asymmetric distributions of preferences, and consequently, asymmetric behavioral patterns.

5.1 Allowing for no recommendation

Consider an extension of the baseline model in which agents have the option of remaining silent—that is, making no recommendation. An M-type individual chooses silence at period t if and only if none of their friends made a recommendation at period $t-1$; in other words, all their friends were silent. An extreme-type agent (either L or R) chooses silence at period t if all of their friends at $t-1$ were either silent or recommended content that is two steps away from the agent's preferred type. This reflects the assumption that agents remain silent when faced with a non-empty set of recommendations only if all such recommendations are two steps removed from their preferred content. This situation never arises for M-types, since any recommendation—whether L or R—is always only one step away from their ideal content.

Let y_t denote the probability that a randomly chosen agent recommends M at period t, and let s_t denote the probability that an agent chooses to remain silent (i.e., makes no recommendation) at period t. By symmetry, the probability of recommending each type of extreme content is then given by $\frac{1-y_t-s_t}{2}$.

The state of society at period t is thus fully described by the vector (y_t, s_t), which satisfies the following constraints:

$$y_t \geq 0, \quad s_t \geq 0, \quad y_t + s_t \leq 1.$$

The dynamical system is now given by

$$y_t = \rho\,(1 - G(1 - y_{t-1})) + (1-\rho)\left(G\left(\frac{1 + y_{t-1} + s_{t-1}}{2}\right) - G\left(\frac{1 - y_{t-1} + s_{t-1}}{2}\right)\right), \tag{15}$$

$$s_t = \rho\,G(s_{t-1}) + (1-\rho)\,G\left(\frac{1 - y_{t-1} + s_{t-1}}{2}\right). \tag{16}$$

Compared to the benchmark model, where the dynamics are governed by equations (2) and (3), the key difference is the introduction of the option to remain silent. This additional behavioral choice results in a new equation, (16), that governs the evolution of silence in the population.

Let (y^*, s^*) be a steady-state equilibrium of the dynamical system defined by (15)–(16). A *no-silence equilibrium* is a steady state in which all agents make a recommendation, that is, $s^* = 0$. By inspecting the system (15)–(16), it is straightforward to verify that the only no-silence equilibrium is given by $(y^*, s^*) = (1, 0)$, that is, we obtain no polarization since each agent recommends a type-M content. Indeed, any other value of $y^* < 1$ combined with $s^* = 0$ would violate equation (16), since the right-hand side would be strictly positive, contradicting the assumption that $s^* = 0$.

A *polarized steady-state equilibrium* is an outcome in which no agent recommends the M-type content, yet a strictly positive share of the population remains active. Formally, this corresponds to $y^* = 0$ and $1 - s^* > 0$.

The next proposition supports the robustness of our earlier findings: namely, that polarization does not arise under high ρ, while full polarization emerges under low ρ.

Proposition 5. *Assume that extreme-type individuals abstain from making a recommendation rather than endorsing content two steps away from their preference and that M-type individuals remain silent at period t if and only if none of their friends made a recommendation at period $t-1$. Then, polarization becomes less likely than in the baseline model, in the following sense:*

(i) The non-polarized equilibrium $(y^*, s^*) = (1, 0)$ is stable if and only if $\rho > \rho_1^{\text{silent}}$, where the threshold ρ_1^{silent} is strictly lower than in the baseline model, that is, $0 \leq \rho_1^{\text{silent}} < \rho_1$.

(ii) In any polarized equilibrium, a strictly positive fraction of the population abstains from recommending content, that is, $s^* > 0$.

(iii) The polarized equilibrium $(y^* = 0, 1 - s^* > 0)$ is stable if and only if $\rho < \rho_0^{\text{silent}}$, where the upper bound ρ_0^{silent} is lower than in the baseline model, that is, $\rho_0^{\text{silent}} < \rho_0$.

The next proposition, which can be viewed as a counterpart to Proposition 2, characterizes how network density affects the degree of polarization.

Proposition 6. *Let $\{p_k(\theta)\}$ be a family of degree distributions satisfying Assumptions 1–3. Then:*

(i) For sparse networks (i.e., θ close to zero), the non-polarized equilibrium is stable, and no non-trivial polarized equilibrium exists.

(ii) There exists a threshold $\widehat{\rho} \in (0,1)$ such that, for all $\rho < \widehat{\rho}$, there exist parameters θ_1 and θ_2 with $0 < \theta_1 < \theta_2 < \overline{\theta}$ satisfying:

$$\theta_1 < \theta < \theta_2 \implies y^*(\theta) = 0 \quad \textit{and} \quad 1 - s^*(\theta) > 0 \quad \textit{(full polarization).}$$

(iii) As $\theta \to \overline{\theta}$, the equilibrium converges to the non-polarized outcome:

$$(y^*(\rho,\theta), s^*(\rho,\theta)) \to (\rho, 0).$$

Proposition 6 shows that allowing for silence preserves the key non-monotonic relationship between network density and polarization established in the baseline model. Specifically, polarization cannot arise in very sparse or very dense networks. In sparse networks (part (i)), individuals have few social contacts and are unlikely to encounter extreme content with enough frequency to adopt or propagate it. As a result, moderate behavior dominates: individuals tend to recommend centrist content, and the society avoids polarization. This reflects a form of behavioral inertia: limited exposure implies limited change. As networks become denser (part (ii)), individuals are more likely to be exposed to content from like-minded peers, enabling reinforcement dynamics and the formation of echo chambers. This facilitates behavioral polarization: a large share of the population recommends only extreme content, while others abstain from engagement. The presence of silence as an option further amplifies this dynamic, as individuals prefer to withdraw rather than recommend content far from their preferences. In very dense networks (part (iii)), the diversity of content exposure increases again, weakening echo chamber effects. Individuals encounter a wider mix of views, including moderate ones, which restores balance and leads to convergence toward a centrist, non-polarized equilibrium. From a behavioral perspective, exposure to diverse viewpoints—enabled by higher connectivity—acts as a moderating force.

Interior equilibria are generally difficult to characterize analytically under an arbitrary degree distribution. To gain sharper insights, we therefore restrict attention to the exponential degree distribution defined in (14), parameterized by the density parameter θ. For this family, we can derive closed-form threshold values that determine the equilibrium structure. Specifically, we define $\rho_0^{\text{silent}}(\theta)$ and $\rho_1^{\text{silent}}(\theta)$ as follows:

$$\rho_0^{\text{silent}}(\theta) := \begin{cases} 0, & \theta \leq 0.75; \\ \text{solves } \rho\dfrac{1}{1-\theta} + (1-\rho)\dfrac{16(1-\theta)}{(2-\theta+\sqrt{4\rho-4\rho\theta+\theta^2})^2} = 1, & \theta > 0.75; \end{cases} \tag{17}$$

$$\rho_1^{\text{silent}}(\theta) := \begin{cases} 0, & \theta \leq 0.5; \\ \dfrac{2\theta-1}{4\theta-(2\theta^2+1)}, & \theta > 0.5. \end{cases} \tag{18}$$

The following proposition refines Proposition 6 by providing a complete characterization of the unique stable equilibrium under the exponential degree distribution.

Proposition 7. *Under the exponential degree distribution* (14) *with density parameter* $\theta \in (0,1)$*, the dynamic system* (15)–(16) *admits a unique stable steady-state equilibrium* (y^*, s^*)*, with the following properties:*

(i) If $\rho \leq \rho_0^{\text{silent}}(\theta) < \rho_0(\theta)$*, the unique stable equilibrium is fully polarized, that is,* $y^* = 0$ *and* $s^* \in (0,1)$*.*

(ii) If $\rho_0^{\text{silent}}(\theta) < \rho < \rho_1^{\text{silent}}(\theta)$*, the unique stable equilibrium is interior, that is,* $y^* \in (0,1)$ *and* $s^* \in (0,1)$*.*

(iii) If $\rho \geq \rho_1^{\text{silent}}(\theta)$*, the unique stable equilibrium is fully centrist with no silence, that is,* $y^* = 1$ *and* $s^* = 0$*.*

(iv) If $\rho < \widehat{\rho}^{\text{silent}} \approx 0.0685$*, then there exist three thresholds* $0 < \theta_1^{\text{silent}} < \theta_2^{\text{silent}} < \theta_3^{\text{silent}} < 1$*, such that the stable equilibrium* $(y^*(\theta,\rho), s^*(\theta,\rho))$ *satisfies:*

$$\begin{aligned}
\theta \leq \theta_1^{\text{silent}} \quad &\Longrightarrow \quad y^* = 1, \quad s^* = 0;\\
\theta_1^{\text{silent}} < \theta < \theta_2^{\text{silent}} \quad &\Longrightarrow \quad 0 < y^* < 1, \quad 0 < s^* < 1;\\
\theta_2^{\text{silent}} \leq \theta \leq \theta_3^{\text{silent}} \quad &\Longrightarrow \quad y^* = 0, \quad 0 < s^* < 1;\\
\theta > \theta_3^{\text{silent}} \quad &\Longrightarrow \quad 0 < y^* < 1, \quad 0 < s^* < 1;\\
\theta \to 1 \quad &\Longrightarrow \quad y^* \to \rho, \quad s^* \to 0.
\end{aligned}$$

Proposition 7 delivers a full characterization of equilibria under an exponential degree distribution. When ρ is low, the population is dominated by ideologically extreme agents (L- and R-types). In such societies, centrists are too rare to sustain the diffusion of M-type content. As a result, the system converges to a polarized equilibrium where $y^* = 0$, and a positive share of agents abstain from communication ($s^* > 0$). In this regime, silence plays a central role: centrist content disappears not because it is rejected, but because there are too few centrists to generate and propagate it. As the centrist share ρ increases beyond the lower threshold $\rho_0^{\text{silent}}(\theta)$, a unique interior equilibrium emerges. Here, all three behavioral modes coexist: some agents recommend M-type content, others promote extreme views, and a share remains silent. This interior regime captures the presence of partial polarization with limited content engagement, reflecting real-world dynamics such as mixed-information environments. For high values of ρ, centrists dominate the population, and the system converges to the fully centrist equilibrium $(y^*, s^*) = (1, 0)$, where all agents recommend moderate content and silence vanishes. Here, centrism becomes self-sustaining: even in dense or clustered networks, the abundance of centrist voices ensures that agents repeatedly encounter moderate content and continue to spread it.

The thresholds $\rho_0^{\text{silent}}(\theta)$ and $\rho_1^{\text{silent}}(\theta)$ are decreasing functions of network density θ, indicating that denser networks reduce the critical mass of centrists needed to avoid polarization. This reflects the fact that exposure to more neighbors increases the likelihood of receiving centrist content, even when centrists are not the majority.

Finally, compared to the baseline model without silence, polarization is less likely in the current extension. This is visible in the lower critical value $\widehat{\rho}^{\text{silent}} \approx 0.0685$ relative to $\widehat{\rho} = 1/9$ in the benchmark case. Allowing silence reduces forced engagement with misaligned content and thus suppresses the reinforcement of extreme views when centrists are scarce.

In sum, the proposition highlights a key structural insight: *the network's capacity to sustain centrism or polarization critically depends on the prevalence of centrist agents and the density of social ties.* Silence acts as a behavioral *filter*: rather than forcing agents to recommend misaligned content, it offers an exit option. This introduces a damping mechanism on polarization and leads to more realistic participation dynamics—capturing real-world political apathy, abstention, or content fatigue.

Figure 3 illustrates Proposition 7 for the exponential degree distribution with the density parameter $\theta = 0.8$. Each of the three panels shows the loci of (s_{t-1}, y_{t-1})-pairs under which, respectively, $y_t = y_{t-1}$ and $s_t = s_{t-1}$. The intersection points of the two loci are steady-state equilibria.[6] Panel 3a shows that, when $\rho > \rho_1^{\text{silent}}(\theta) = \frac{15}{23}$, the unique stable equilibrium is the non-polarized equilibrium ($y^* = 1$). Panel 3b shows that, when $\rho < \rho_0^{\text{silent}}(\theta) \approx 0.0522$, the unique stable equilibrium is the polarized equilibrium ($y^* = 0$). Finally, panel 3c shows that, when $\rho_0^{\text{silent}}(\theta) < \rho < \rho_1^{\text{silent}}(\theta)$, the unique stable equilibrium is the interior equilibrium ($0 < y^* < 1$).

Figure 3 also provides phase diagrams which allow one establish qualitatively the stability of the non-polarized equilibrium on panel 3a, the polarized equilibrium on panel 3b, and the interior equilibrium on panel 3c. The directions of the dynamics of s_t, indicated by horizontal arrows, and those of the dynamics of y_t, indicated by vertical arrows, are determined from, respectively, equations (35) and (37) in the Appendix.

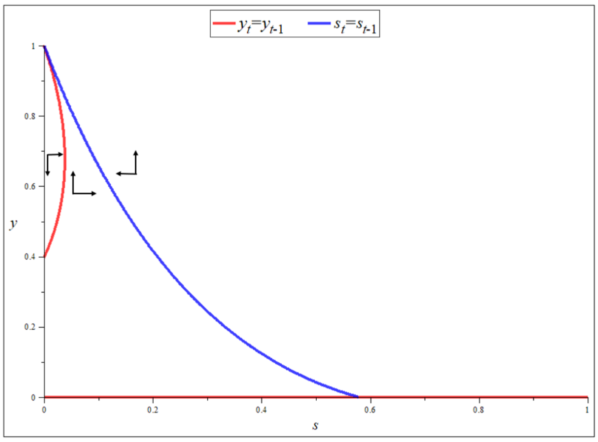
Proposition 7 and Figure 3 jointly demonstrate that introducing the option of silence narrows the parameter space in which polarization can occur. Compared to the baseline model (Proposition 3), the critical threshold $\widehat{\rho}^{\text{silent}} \approx 0.0685$ is lower than its baseline counterpart $\widehat{\rho} = 1/9 \approx 0.1111$. This means that silence acts as a moderating force: it reduces the likelihood of individuals recommending extreme content when no moderate content is available.

5.2 Allowing for asymmetry

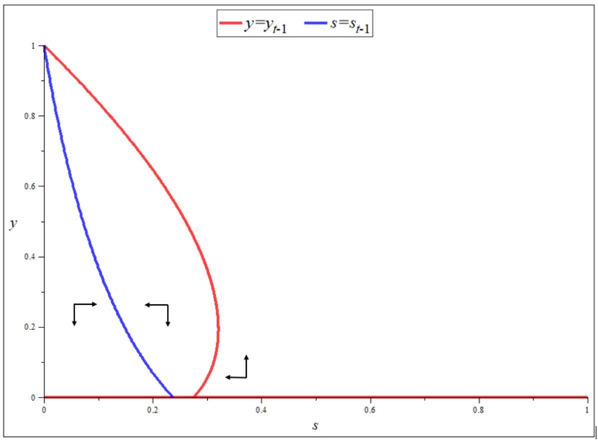
Our second robustness check relaxes the assumption of symmetry in the distribution of extreme agents. Specifically, let $\epsilon \in \left(-\frac{1-\rho}{2}, \frac{1-\rho}{2}\right)$ denote the *asymmetry parameter*, where a positive (negative) value of ϵ captures skewness toward right- (left-) oriented individuals. The resulting distribution of agent types is given by:

Type	L	M	R
Fraction	$\frac{1-\rho}{2} - \epsilon$	ρ	$\frac{1-\rho}{2} + \epsilon$

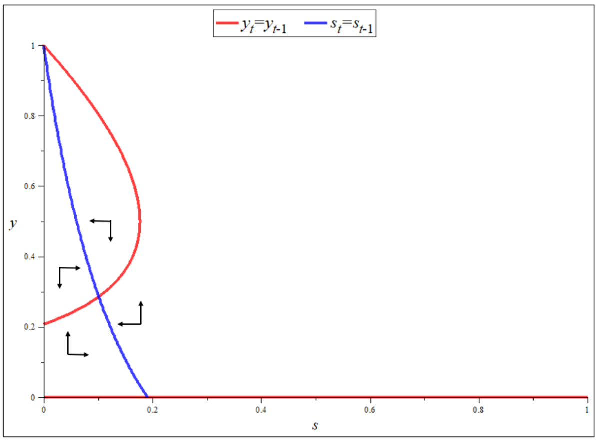
[6] Technically, there is one more steady state—the "full-silence" steady state, in which $s^* = 1$. However, we show in Online Appendix D that this steady state is never stable, therefore we disregard it as uninteresting.



(a) The case of the non-polarized stable equilibrium: $\rho = 0.75 > \rho_1^{\text{silent}}(\theta) = \frac{15}{23}$



(b) The case of the polarized stable equilibrium: $\rho = 0.04 < \rho_0^{\text{silent}}(\theta) \approx 0.0522$



(c) The case of the interior stable equilibrium: $\rho = 0.2 \in \left(0.0522, \frac{15}{23}\right)$

Figure 3: Equilibrium characterization under exponential distribution with $\theta = 0.8$.

As a result, the fractions of L-type and R-type agents are no longer symmetric around the centrist share ρ. When $\epsilon = 0$, we recover the benchmark model with a symmetric distribution.

In what follows, we focus, without loss of generality, on the case $\epsilon > 0$, which corresponds to a right-skewed preference distribution in which R-type agents are more prevalent than L-types. The case $\epsilon < 0$ (left-skewed) is entirely analogous, as it constitutes the mirror image of the right-skewed scenario. Let ℓ_t and r_t denote the probabilities that a randomly selected agent recommends L-type and R-type content, respectively, at time t. These two state variables fully characterize the state of society at period t. The evolution of the system is governed by the following law of motion:

$$\ell_t = \left(\frac{1-\rho}{2} - \epsilon\right)(1 - G(1-\ell_{t-1})) + \rho\,\frac{\ell_{t-1}}{\ell_{t-1}+r_{t-1}}\,G(\ell_{t-1}+r_{t-1}) + \left(\frac{1-\rho}{2} + \epsilon\right)G(\ell_{t-1}), \quad (19)$$

$$r_t = \left(\frac{1-\rho}{2} + \epsilon\right)(1 - G(1-r_{t-1})) + \rho\,\frac{r_{t-1}}{\ell_{t-1}+r_{t-1}}\,G(\ell_{t-1}+r_{t-1}) + \left(\frac{1-\rho}{2} - \epsilon\right)G(r_{t-1}). \quad (20)$$

The second terms of the right-hand side of equations (19) and (20), given respectively by

$$\rho\,\frac{\ell_{t-1}}{\ell_{t-1}+r_{t-1}}\,G(\ell_{t-1}+r_{t-1}) \quad \text{and} \quad \rho\,\frac{r_{t-1}}{\ell_{t-1}+r_{t-1}}\,G(\ell_{t-1}+r_{t-1}),$$

reflect the following behavioral convention: if each friend of an M-type agent recommends either L- or R-type content (and none recommend M-type), then the agent is indifferent and randomizes uniformly across the observed recommendations. Specifically, if an M-type agent has k friends, $j \leq k$ of whom recommend L-type content while the remaining $k-j$ recommend R-type, then the agent adopts the L-type recommendation with probability j/k.

The expression

$$\frac{\ell_{t-1}}{\ell_{t-1}+r_{t-1}}\,G(\ell_{t-1}+r_{t-1})$$

thus represents the conditional probability that an M-type agent recommends L-type content, which can be derived algebraically as follows:

$$\begin{aligned}
\frac{\ell_{t-1}}{\ell_{t-1}+r_{t-1}} \sum_k p_k (\ell_{t-1}+r_{t-1})^k &= \ell_{t-1} \sum_k p_k (\ell_{t-1}+r_{t-1})^{k-1} \\
&= \ell_{t-1} \sum_k p_k \sum_{j=0}^{k-1} \ell_{t-1}^j r_{t-1}^{k-1-j} \binom{k-1}{j} \\
&= \sum_k p_k \sum_{j=1}^{k} \ell_{t-1}^j r_{t-1}^{k-j} \binom{k-1}{j-1} \\
&= \sum_k p_k \sum_{j=1}^{k} \ell_{t-1}^j r_{t-1}^{k-j} \frac{j}{k} \binom{k}{j},
\end{aligned}$$

where the last line gives the full expression for the probability that an M-type agent randomly adopts an L-type recommendation when exposed only to extreme content.

As in Section 3, we define a *non-polarized equilibrium* as a steady state in which no agent recommends extreme content (i.e., $\ell^* = r^* = 0$), and a *polarized equilibrium* as one in which no agent recommends M-type content.

Proposition 8. *We obtain the following results for a right-skewed preference distribution ($\epsilon > 0$):*

(i) The non-polarized equilibrium is stable if and only if

$$\rho > \rho_1(\epsilon) := \frac{\mu_1 + p_1 - 2}{\mu_1 - p_1} + 2\epsilon.$$

The threshold level $\rho_1(\epsilon)$ is strictly higher than the corresponding threshold ρ_1 in the baseline model, and increases with the degree of asymmetry ϵ. In the limit, we have

$$\lim_{\epsilon \to 0} \rho_1(\epsilon) = \rho_1.$$

(ii) There exists a threshold $\rho_0(\epsilon) < \rho_1(\epsilon)$ such that the polarized equilibrium is stable if and only if

$$\rho < \rho_0(\epsilon).$$

If p_2 is not too large, then $\rho_0(\epsilon) < \rho_0$, where ρ_0 is the corresponding threshold in the baseline model. Moreover, $\rho_0(\epsilon)$ decreases with the degree of asymmetry ϵ, and

$$\lim_{\epsilon \to 0} \rho_0(\epsilon) = \rho_0.$$

(iii) When $\rho \in (\rho_0(\epsilon), \rho_1(\epsilon))$, there exists a unique interior stable equilibrium $(\ell^*(\epsilon), r^*(\epsilon))$.

Proposition 8 characterizes how the steady-state equilibrium of the system depends on the centrist share ρ and the degree of asymmetry ϵ in the population's ideological distribution. It shows that even a small departure from symmetry in the distribution of extreme agents significantly alters the long-run behavior of the system. When the population becomes slightly more skewed toward R-type individuals (i.e., $\epsilon > 0$), both the range of parameters sustaining full centrism (non-polarized equilibrium) and those sustaining full extremism (polarized equilibrium) become narrower.

Part (i) establishes that a non-polarized equilibrium—where only M-type content is recommended—is stable if and only if the share of centrist agents exceeds the threshold $\rho_1(\epsilon)$. This threshold is increasing in ϵ: when the distribution becomes more skewed (e.g., more R-type agents than L-types), centrism becomes harder to sustain. Intuitively, the presence of more extremists on one side distorts exposure and recommendation dynamics in favor of that extreme, requiring a higher critical mass of centrists to counterbalance this effect and restore ideological consensus. Part (ii) characterizes the polarized equilibrium—where no M-type content is recommended. It is stable if and only if the centrist share ρ falls below the threshold $\rho_0(\epsilon)$. Asymmetry lowers this threshold, meaning that even fewer centrists are needed to destabilize full polarization. The reason is that in a skewed society, the dominant extreme becomes more effective in amplifying its own content, even in the presence of a modest centrist minority. Hence, asymmetry makes polarized equilibria more fragile with respect to small increases in ρ. Part (iii) shows that when the centrist share lies between the two thresholds, $\rho \in (\rho_0(\epsilon), \rho_1(\epsilon))$, the system converges to an interior equilibrium, where both types of extreme content coexist and receive positive recommendation shares, but M-type content also circulates. This reflects partial polarization, with the network fragmenting into ideological subgroups.

Overall, the proposition demonstrates that the range of centrist shares that supports full polarization or full centrism shrinks under asymmetry, and that a larger interior region emerges. That is, as the distribution becomes ideologically unbalanced (e.g., more R-types than L-types), the system is less likely to settle in a clean consensus (centrism) or total ideological dominance (polarization). Instead, it more often stabilizes in mixed states with partial ideological coexistence and content fragmentation. From a modeling standpoint, this illustrates how the structure of the population—especially imbalances across ideological types—shapes the qualitative nature of the communication equilibrium, even when preferences themselves remain fixed.

6 Concluding remarks

We have developed a simple yet powerful model that illustrates how the structure of social networks influences the diffusion of political content and the degree of ideological polarization. Despite its tractability, the model yields rich equilibrium behavior and offers insights into when connectivity fosters consensus versus when it amplifies division.

First, we show that centrist content benefits from a pairwise advantage—since no agent is more than one ideological step away from the center—which leads to amplification of moderate views when centrists are sufficiently prevalent. However, when extremist preferences dominate, sufficiently dense networks can reinforce polarization by enabling like-minded individuals to encounter and forward extreme content. This mechanism may help explain the rise in political polarization observed in settings like the United States. Second, we establish that the relationship between network density and polarization is non-monotonic: intermediate levels of connectivity are most likely to generate polarization, while very sparse or very dense networks tend to favor moderation. This finding helps reconcile conflicting empirical results on the impact of social media on political attitudes. Third, we find that partial polarization is less likely to arise in regular networks compared to those with a more dispersed degree distribution. This suggests that heterogeneity in connectivity—i.e., the presence of highly connected individuals or "influencers"—can foster fragmentation and ideological sorting, even when average exposure remains unchanged. We also explore two robustness extensions. In the first, agents are allowed to abstain from forwarding ideologically distant content, which acts as a moderating force but does not eliminate polarization under dispersed preferences. In the second, we introduce asymmetry in the distribution of ideological types and find that both full- and no-polarization outcomes become less robust, making partial polarization more likely.

A promising direction for future work is to endogenize the network formation process. Agents could strategically choose their connections based on ideological proximity or content alignment, leading to endogenous echo chambers or ideological segregation. Such an extension would provide a deeper understanding of how social fragmentation and polarization co-evolve but would be less tractable..

Our results also suggest important policy implications. Efforts to reduce polarization by increasing exposure to diverse views may backfire if implemented in networks with intermediate density. Instead, interventions should consider both the ideological composition of the population and the structure of communication. Policies that promote broad, weakly tied connections—rather than highly clustered ones—may be more effective in counteracting polarization. Moreover, limiting the influence of highly connected individuals or amplifying centrist voices could help stabilize moderate discourse in politically fragmented societies.

Proofs of all results in the main text

Proof of Proposition 1. Let us start with the following lemma, which will be useful later in proving the uniqueness of a stable equilibrium.
Lemma 1.

(i) The function $f(y,\rho)$ defined by (3) is strictly increasing for all $y \in [0,1]$.

(ii) Only one of the three mutually exclusive statements holds:

(iia) $f_{yy} > 0\ \forall y \in (0,1)$,

(iib) $\exists! \widehat{y} \in (0,1)$, *such that* $f_{yy} \lesseqgtr 0 \Longleftrightarrow y \lesseqgtr \widehat{y}$,

(iic) $f_{yy} < 0\ \forall y \in (0,1)$.

Proof of Lemma 1. (i) By differentiating (3) with respect to y, we obtain:

$$f_y(y,\rho) = \rho G'(1-y) + (1-\rho)\left[\frac{1}{2}G'\left(\frac{1+y}{2}\right) + \frac{1}{2}G'\left(\frac{1-y}{2}\right)\right]. \tag{21}$$

Combining (21) with (**??**) proves (i).

(ii) Equation (21) implies that $f_y(y,\rho,)$ is convex with respect to y, since it is a convex combination of three convex functions of y, that is, $G'(1-y)$, $G'\left(\frac{1+y}{2}\right)$, and $G'\left(\frac{1-y}{2}\right)$. Hence, $f_{yyy}(y,\rho,) > 0$, $\forall y \in (0,1)$. Thus, we have:

$$f_{yy}(0,\rho) \geq 0 \Longrightarrow \text{alternative } (ii.a);$$
$$f_{yy}(0,\rho) < 0 < f_{yy}(1,\rho) \Longrightarrow \text{alternative } (ii.b);$$
$$f_{yy}(1,\rho) \leq 0 \Longrightarrow \text{alternative } (ii.c).$$

This proves (ii) and completes the proof. □

We first prove that the stable steady state is unique. From the definition (3) of the f-function, one can see that both $y=0$ and $y=1$ are solutions to the steady state condition (5); hence, they are both steady states.

If the alternative $(ii.a)$ from Lemma 1 prevails, the f-function lies below the 45°-degree line for all $y \in (0,1)$. Hence, no interior solutions exist. In this case, $f_y(0,\rho) < 1 < f_y(1,\rho)$, and the unique stable steady state is the full polarization solution $y^* = 0$.

Similarly, if the alternative $(ii.c)$ from Lemma 1 takes place, the f-function lies above the 45°-degree line for all $y \in (0,1)$. Again, no interior solutions exist. In this case, $f_y(0,\rho) > 1 > f_y(1,\rho)$, and the unique stable steady state is the no polarization solution, $y^* = 1$.

If the alternative $(ii.b)$ from Lemma 1 prevails, and $f_y(0,\rho,\{p_k\}) < 1$, then the full polarization steady state is stable, that is,

$$f_{yy}(1,\rho) > 0 \Longrightarrow f_y(1,\rho) > 1.$$

Finally, if the alternative $(ii.b)$ from Lemma 1 takes place, and $f_y(0, \rho) > 1$, then the full polarization steady state is unstable and so is the no-polarization state, since

$$f_{yy}(1, \rho, \{p_k\}) > 0 \Longrightarrow f_y(1, \rho, \{p_k\}) > 1.$$

Hence, $f(y, \rho, \{p_k\}) - y > 0$ in the vicinity of $y = 0$, and $f(y, \rho, \{p_k\}) - y < 0$ in the vicinity of $y = 1$. By the intermediate value theorem, there must be $y^* \in (0, 1)$ which solves the steady state condition (5). Furthermore, as $f(\cdot, \rho, \{p_k\})$ only has one inflection point, the interior solution y^* is unique. Finally, we have:

$$f_y(0, \rho, \{p_k\}) > 1 \Longleftrightarrow \frac{\partial}{\partial y}\left[f(y, \rho, \{p_k\}) - y\right]\bigg|_{y=0} > 0;$$

$$f_y(1, \rho, \{p_k\}) > 1 \Longleftrightarrow \frac{\partial}{\partial y}\left[f(y, \rho, \{p_k\}) - y\right]\bigg|_{y=1} > 0.$$

Hence, it must be that

$$\frac{\partial}{\partial y}\left[f(y, \rho, \{p_k\}) - y\right]\bigg|_{y=y^*} < 0 \quad \Longrightarrow \quad f_y(y^*, \rho, \{p_k\}) < 1.$$

Thus, the interior solution y^* is the unique stable solution.

It remains to prove that $y^*_\rho > 0$ when $\rho_0 < \rho < \rho_1$. Differentiating both sides of the steady state condition (5) with respect to ρ, we obtain:

$$y^*_\rho = \frac{f_\rho(y, \rho)}{1 - f_y(y, \rho)}\bigg|_{y=y^*}.$$

Since $f_y(y^*, \rho) < 1$, due to the stability of the interior steady state, we have:

$$\text{sign}\left\{y^*_\rho\right\} = \text{sign}\left\{f_\rho(y, \rho)\right\}\big|_{y=y^*}.$$

From (3), and from $G(1) = 1$,

$$f_\rho(y, \rho) = \left[G(1) - G\left(\frac{1+y}{2}\right)\right] - \left[G(1-y) - G\left(\frac{1-y}{2}\right)\right].$$

Since $1 - \frac{1+y}{2} = (1-y) - \frac{1-y}{2} = \frac{1-y}{2}$, we have $G(1) - G\left(\frac{1+y}{2}\right) > G(1-y) - G\left(\frac{1-y}{2}\right)$, due to convexity of $G(\cdot)$. Hence, $f_\rho(y, \rho) > 0 \Longrightarrow y^*_\rho > 0$. This completes the proof. $\square$

Proof of Proposition 2

We will need the following Lemma.

Lemma 2. *Under Assumptions 1-3, $\rho_1(\theta)$ increases with θ, and*

$$\lim_{\theta\to 0}\rho_1(\theta)=0 \tag{22}$$

$$\lim_{\theta\to\bar{\theta}}\rho_1(\theta)=1. \tag{23}$$

Proof of Lemma 2. From Assumption 2, $\lim_{\theta\to\bar{\theta}}\mu_1(\theta)\to\infty$, hence (23) follows immediately from (7). Next, from Assumption 3, we get, under $\theta\to 0$

$$\mu_1(\theta)=1+\lambda\theta+o(\theta)\Longrightarrow\rho_1(\theta)=\frac{o(\theta)}{\theta},$$

which implies (22).

To prove that $\rho_1'(\theta)>0$, let us first rewrite (7) as follows:

$$\rho_1(\theta)=\frac{\frac{\mu_1(\theta)-1}{1-p_1(\theta)}-1}{\frac{\mu_1(\theta)-1}{1-p_1(\theta)}+1},$$

where the RHS is a monotone transformation of $\frac{\mu_1(\theta)-1}{1-p_1(\theta)}$, hence

$$\rho_1'(\theta)>0\Longleftrightarrow\frac{d}{d\theta}\left[\frac{\mu_1(\theta)-1}{1-p_1(\theta)}\right]>0\Longleftrightarrow\frac{\mu_1'(\theta)}{\mu_1(\theta)-1}+\frac{p_1'(\theta)}{1-p_1(\theta)}>0.$$

Let us rewrite the expression $\frac{\mu_1'(\theta)}{\mu_1(\theta)-1}+\frac{p_1'(\theta)}{1-p_1(\theta)}$, which we need to sign, as follows:

$$\frac{\mu_1'(\theta)}{\mu_1(\theta)-1}+\frac{p_1'(\theta)}{1-p_1(\theta)}=\sum_{k=1}^{\infty}\frac{p_k'(\theta)}{p_k(\theta)}a_k(\theta), \tag{24}$$

where the coefficinets $a_k(\theta)$ are defined by

$$a_k(\theta):=\begin{cases}0, & k=1;\\ \left(\frac{k-1}{\mu_1(\theta)-1}-\frac{1}{1-p_1(\theta)}\right)p_k(\theta), & k\geq 2.\end{cases} \tag{25}$$

Representation (24) – (25) is readily verified using the following identities:

$$\frac{\mu_1'(\theta)}{\mu_1(\theta)-1}=\frac{\sum_{k=1}^{\infty}(k-1)p_k'(\theta)}{\sum_{k=1}^{\infty}(k-1)p_k(\theta)},$$

$$\frac{p_1'(\theta)}{1-p_1(\theta)}=-\frac{\sum_{k=1}^{\infty}(1-\delta_{1k})p_k'(\theta)}{\sum_{k=1}^{\infty}(1-\delta_{1k})p_k(\theta)},$$

where δ_{1k} is the Kronecker's delta. The sequence $\{a_k(\theta)\}$ defined by (25) satisfies the following two properties. First,

$$\sum_{k=1}^{\infty} a_k(\theta) = 0.$$

Second, as k increases, $a_k(\theta)$ changes sign only once, from "$-$" to "$+$":

$$\exists! k_0 > 2 : a_k(\theta) > 0 \Longleftrightarrow k > k_0.$$

Let us define:

$$A := -\sum_{k \le k_0} a_k(\theta) = \sum_{\ell > k_0} a_\ell(\theta) > 0$$

and restate (24) as follows:

$$\begin{aligned}
\frac{\mu_1'(\theta)}{\mu_1(\theta)-1} + \frac{p_1'(\theta)}{1-p_1(\theta)} &= \sum_{\ell > k_0} \frac{p_\ell'(\theta)}{p_\ell(\theta)} a_\ell(\theta) - \sum_{k \le k_0} \frac{p_k'(\theta)}{p_k(\theta)} \left(-a_k(\theta)\right) \\
&= A \left[\sum_{\ell > k_0} \frac{p_\ell'(\theta)}{p_\ell(\theta)} \frac{a_\ell(\theta)}{A} - \sum_{k \le k_0} \frac{p_k'(\theta)}{p_k(\theta)} \left(-\frac{a_k(\theta)}{A} \right) \right] \\
&> A \left[\inf_{\ell > k_0} \left\{ \frac{p_\ell'(\theta)}{p_\ell(\theta)} \right\} - \sup_{k \le k_0} \left\{ \frac{p_k'(\theta)}{p_k(\theta)} \right\} \right] > 0,
\end{aligned}$$

where the last inequality follows from Assumption 1. Indeed, as θ respects the MLR ordering,

$$\ell > k \Longrightarrow \frac{p_\ell'(\theta)}{p_\ell(\theta)} > \frac{p_k'(\theta)}{p_k(\theta)} \Longrightarrow \inf_{\ell > k_0} \left\{ \frac{p_\ell'(\theta)}{p_\ell(\theta)} \right\} > \sup_{k \le k_0} \left\{ \frac{p_k'(\theta)}{p_k(\theta)} \right\}.$$

This proves that $\rho'(\theta) > 0$, and completes the proof. $\square$

We now proceed with the proof of Proposition 2.

(i) From Lemma 2 and the intermediate value theorem, the equation $\rho_1(\theta) = \rho$ has the unique solution $\theta_1(\rho) \in \left(0, \overline{\theta}\right)$ for every $\rho \in (0,1)$. Condition (10) guarantees that $y^* \to \rho$ as $\theta \to \overline{\theta}$. From Lemma B1 in the Online Appendix B, $\rho_0(\theta) < \rho_1(\theta)$; hence, $\lim_{\theta \to 0} \rho_0(\theta) = 0$. Also, from (6), $\lim_{\theta \to \overline{\theta}} \rho_0(\theta) = 0$ since $\lim_{\theta \to \overline{\theta}} \mu_1(\theta) = \infty$. Define

$$\widehat{\rho} := \sup_{\theta \in \left(0, \overline{\theta}\right)} \rho_0(\theta) \in (0,1).$$

If $\rho \in (0, \widehat{\rho})$, the equation $\rho_0(\theta) = \rho$ has at least two distinct roots, $\theta_2(\rho)$ and $\theta_3(\rho) > \theta_2(\rho)$. Furthermore, $\theta_2(\rho) > \theta_1(\rho)$, as $\rho_1(\theta) > \rho_0(\theta)$. Hence, if $\rho < \widehat{\rho}$, in the simplest case when the equation $\rho_0(\theta) = \rho$ only has two roots (which is the case, for example, when the distribution is exponential), we get:

$$\theta \le \theta_1(\rho) \Longrightarrow y^*(\theta, \rho) = 1 \text{ (no polarization);}$$

$$\theta \in (\theta_1(\rho), \theta_2(\rho)) \Longrightarrow y^*(\theta, \rho) \in (0,1) \;\; \text{(partial polarization);}$$

$$\theta \in [\theta_2(\rho), \theta_3(\rho)] \Longrightarrow y^*(\theta, \rho) = 0 \; \text{(full polarization);}$$

$$\theta > \theta_3(\rho) \Longrightarrow y^*(\theta, \rho) \in (0,1) \;\; \text{(partial polarization);}$$

$$\theta \to \overline{\theta} \Longrightarrow y^*(\theta, \rho) \to \rho.$$

This proves (i).[7]

(ii) Since $\rho(\theta)$ increases from 0 to 1 as θ increases from 0 to $\overline{\theta}$, the equation $\rho_1(\theta) = \rho$ has the unique solution $\widehat{\theta}(\rho)$ for each $\rho \in (0,1)$, and

$$\rho \gtreqless \rho_1(\theta) \Longleftrightarrow \theta \lesseqgtr \widehat{\theta}(\rho).$$

Thus, (ii) follows from Proposition 1.

(iii) The result follows immediately if we set $\theta \to \overline{\theta}$ in (5) and notice that, from Assumption 2, $\lim_{\theta \to \overline{\theta}} G(x, \theta) = 0$. This proves (iii) and completes the proof. □

Proof of Proposition 3

The generating function of the exponential distribution (14) is given by:

$$G(x) = \frac{(1-\theta)x}{1-\theta x}.$$

Plugging this generating function into (3) and solving the steady-state condition (5), it is readily verified that the closed form for the stable steady-state equilibrium $y^*(\theta, \rho)$ is given by

$$y^*(\theta, \rho) = \max\{0, \min\{1, \widetilde{y}(\theta, \rho)\}\},$$

where

$$\widetilde{y}(\theta, \rho) := \frac{\sqrt{(1-\rho-2\theta)^2 + 16\rho(1-\theta)} - (1-\rho)}{2\theta}. \tag{26}$$

The function $\widetilde{y}(\theta, \rho)$ defined by (26) is quasiconvex with respect to θ over $(0,1)$, as the equation $\widetilde{y}(\theta, \rho) = y$ cannot have more than two solutions in $(0,1)$ for any $y \in \mathbb{R}$. Indeed, using (26), the equation $\widetilde{y}(\theta, \rho) = y$ can be restated as

$$(1-\rho-2\theta)^2 + 16\rho(1-\theta) - (2y\theta + 1 - \rho)^2 = 0,$$

which is a quadratic equation in θ, hence it has at most two real solutions. Furthermore, by setting

[7] If the equation $\rho_0(\theta) = \rho$ has more than two roots, there will be multiple switches between partial polarization and full polarization as θ grows. We have ruled out this opportunity for the special case of the exponential distribution (see Proposition 3) but not in general.

$\theta \to 0$ and $\theta \to 1$ in the RHS of (26), we get:

$$\lim_{\theta \to 0} \widetilde{y}(\theta, \rho) = +\infty, \qquad \lim_{\theta \to 1} \widetilde{y}(\theta, \rho) = \rho < +\infty.$$

Hence, $\widetilde{y}(\theta, \rho)$ is U-shaped with respect to θ over $(0, 1)$ if and only if

$$\lim_{\theta \to 1} \widetilde{y}_\theta(\theta, \rho) > 0,$$

and is decreasing otherwise.

We now prove statements (i)-(iii).

(i) Equation $\widetilde{y}(\theta, \rho) = 0$, which, from (26), is equivalent to the quadratic equation

$$\theta^2 - (1 + 3\rho)\,\theta + 4\rho = 0,$$

has two distinct roots, $0 < \theta_1 < \theta_2 < 1$, if and only if the discriminant is positive:

$$D = (1 + 3\rho)^2 - 16\rho > 0 \Longleftrightarrow \rho < \frac{1}{9}.$$

This proves (i).

(ii) If $\frac{1}{9} < \rho < \sqrt{5} - 2$, we have $\widetilde{y}(\theta, \rho) > 0\ \forall\theta(0, 1)$. Also, we have:

$$\lim_{\theta \to 0} \widetilde{y}_\theta(\theta, \rho) = +\infty, \quad \lim_{\theta \to 1} \widetilde{y}_\theta(\theta, \rho) = 1 - \rho - \frac{4\rho}{1 + \rho} > 0.$$

Thus, $\widetilde{y}(\theta, \rho)$ has an interior minimizer w.r.t. θ, and hence it varies non-monotonically. This proves (ii).

(iii) If $\rho \geq \sqrt{5} - 2$, we have $\widetilde{y}(\theta, \rho) > 0\ \forall\theta(0, 1)$. Also, we have:

$$\lim_{\theta \to 1} \widetilde{y}_\theta(\theta, \rho) = 1 - \rho - \frac{4\rho}{1 + \rho} \leq 0.$$

Thus, $\widetilde{y}(\theta, \rho)$ monotonically decreases from 1 to ρ. This proves (iii) and completes the proof. □

Proof of Proposition 4

First, consider the impact of a mean preserving spread on ρ_0. Assume first that p_1 remains unchanged as a result of the mean-preserving spread. Then, as the expression $k \cdot (1/2)^{k-1}$ is a convex function of k for all $k \geq 2$, $\psi = \mathbb{E}\left[\widetilde{k} \cdot (1/2)^{\widetilde{k}-1}\right]$ increases in response to a mean-preserving spread. From (6), the threshold value ρ_0 decreases with respect to ψ. We conclude that ρ_0 falls in response to a mean-preserving spread when p_1 remains unchanged. By the standard continuity argument, this remains to be true if p_1 increases not too much. from (7), ρ_1 does not change either.

Consider now the impact of a mean-preserving spread on ρ_1. From (7), ρ_1 increases w.r.t. p_1, hence ρ_1 weakly increases in response to a mean preserving spread which keeps μ_1 unchanged and leads to a weak increase in p_1. This completes the proof. $\square$

Proof of Proposition 5

From (15)-(16), the steady-state conditions are:

$$y = \rho\left(1 - G(1-y)\right) + (1-\rho)\left(G\left(\frac{1+y+s}{2}\right) - G\left(\frac{1-y+s}{2}\right)\right); \tag{27}$$

$$s = \rho\, G(s) + (1-\rho)\, G\left(\frac{1-y+s}{2}\right). \tag{28}$$

The Jacobian evaluated at a steady state (y^*, s^*) is given by:

$$\mathbf{J}(y^*, s^*) = \begin{pmatrix} \rho G'(1-y^*) + \frac{1-\rho}{2}\left(G'\left(\frac{1+y^*+s^*}{2}\right) + G'\left(\frac{1-y^*+s^*}{2}\right)\right) & \frac{1-\rho}{2}\left(G'\left(\frac{1+y^*+s^*}{2}\right) - G'\left(\frac{1-y^*+s^*}{2}\right)\right) \\ -\frac{1-\rho}{2} G'\left(\frac{1-y^*+s^*}{2}\right) & \rho\, G'(s^*) + \frac{1-\rho}{2}\, G'\left(\frac{1-y^*+s^*}{2}\right) \end{pmatrix}. \tag{29}$$

(i) The no-polarization case is an equilibrium (y^*, s^*), which satisfies $y^* + s^* = 1$ (everyone either recommends M-type content or is inactive) and $1 - s^* > 0$ (the fraction of active individuals is positive). Using the steady-state conditions, it is readily verified that there is only one such equilibrium given by:

$$(y^*, s^*) = (1, 0).$$

Plugging $(y^*, s^*) = (1, 0)$ into the expression (29) for the Jacobian evaluated at the steady state, we get the Jacobian $\mathbf{J}(1,0)$ evaluated at the non-polarized equilibrium:

$$\mathbf{J}(1,0) = \begin{pmatrix} \left(\rho + \frac{1-\rho}{2}\right) p_1 + \frac{1-\rho}{2}\mu_1 & \frac{1-\rho}{2}\left(\mu_1 - p_1\right) \\ -\frac{1-\rho}{2} p_1 & \left(\rho + \frac{1-\rho}{2}\right) p_1 \end{pmatrix}.$$

The trace and the determinant of $\mathbf{J}(1,0)$ are given, respectively, by:

$$\mathrm{tr}\left(\mathbf{J}(1,0)\right) = \rho p_1 + \frac{1-\rho}{2}\mu_1 + p_1;$$

$$\det\left(\mathbf{J}(1,0)\right) = \left(\rho p_1 + \frac{1-\rho}{2}\mu_1\right) p_1.$$

From the Vieta theorem, the eigenvalues of $\mathbf{J}(1,0)$ are given by:

$$\lambda_1 = \rho p_1 + \frac{1-\rho}{2}\mu_1;$$

$$\lambda_2 = p_1.$$

The no-polarization equilibrium is stable iff $|\lambda_i| < 1$ for $i = 1, 2$, which in turn holds iff the following condition holds:

$$\rho > \rho_1^{\text{silent}} := \begin{cases} 0, & \mu_1 \leq 2; \\ \frac{\frac{\mu_1}{2}-1}{\frac{\mu_1}{2}-p_1}, & \mu_1 > 2. \end{cases} \tag{30}$$

It is readily verified that $0 \leq \rho_1^{\text{silent}} < \rho_1$, where ρ_1 is the lower bound from the baseline model given by (7). This proves (i).

(ii) The steady-state condition obtained from (15) holds for $y^* = 0$ and for every $s^* \in [0, 1]$, hence it can be disregarded. The steady-state condition obtained from (16) becomes under $y^* = 0$:

$$s = \rho G(s) + (1 - \rho) G\left(\frac{1+s}{2}\right). \tag{31}$$

Because the RHS of (31) is an increasing and convex function of s, positive at $s = 0$ and equal to one at $s = 1$, (31) has a non-trivial solution $s^* < 1$ if and only if the slope of the RHS exceeds one:

$$\left(\rho + \frac{1-\rho}{2}\right) G'(1) > 1,$$

or, equivalently,

$$\mu_1 > \frac{2}{1+\rho}.$$

As $G\left(\frac{1}{2}\right) > 0$, it is always true that, whenever s^* is well defined, it is strictly positive. This proves (ii).

(iii) Evaluating the Jacobian given by (29) at $(0, s^*)$, we get

$$\begin{pmatrix} \rho G'(1) + (1-\rho) G'\left(\frac{1+s^*}{2}\right) & 0 \\ -\frac{1-\rho}{2} G'\left(\frac{1+s^*}{2}\right) & \rho G'(s^*) + \frac{1-\rho}{2} G'\left(\frac{1+s^*}{2}\right) \end{pmatrix},$$

its eigenvalues being $\lambda_1 = \rho G'(1) + (1-\rho) G'\left(\frac{1+s^*}{2}\right)$ and $\lambda_2 = \rho G'(s^*) + \frac{1-\rho}{2} G'\left(\frac{1+s^*}{2}\right)$. It is readily verified that $\lambda_1 > \lambda_2 > 0$. Hence, the stability condition for the polarized equilibrium is $\lambda_1 < 1$. This holds if ρ is sufficiently small and $G'\left(\frac{1+s^*}{2}\right) < 1$ (which occurs if, for example, the network is sufficiently dense, hence $G'(\cdot)$ is close to zero almost everywhere). If both these conditions hold, we get

$$\text{polarized equilibrium is stable} \Longleftrightarrow \rho < \rho_0^{\text{silent}} := \frac{1 - G'\left(\frac{1+s^*}{2}\right)}{G'(1) - G'\left(\frac{1+s^*}{2}\right)}.$$

Observe that the threshold fraction ρ_0^{silent} of middle-oriented individual for the polarized equilibrium to be stable is lower in the case where silence is an option than in the baseline case. Indeed, for the baseline case the threshold ρ_0 is given by (6), and we get:

$$\rho_0^{\text{silent}} = \frac{1 - G'\left(\frac{1+s^*}{2}\right)}{G'(1) - G'\left(\frac{1+s^*}{2}\right)} < \frac{1 - G'\left(\frac{1}{2}\right)}{G'(1) - G'\left(\frac{1}{2}\right)} = \rho_0.$$

This completes the proof. $\square$

Proof of Proposition 6

(i) The polarized equilibrium fails to exist if

$$\mu(\theta) \leq \frac{2}{1+\rho} \Longrightarrow \theta \leq \mu^{-1}\left(\frac{2}{1+\rho}\right).$$

This proves non-existence of a polarized equilibrium for very sparce networks. The stability of the non-polarized equilibrium follows from Proposition 5(i) and Lemma 2. This proves (i)

(ii) We proceed in three steps.

Step 1: defining $\widehat{\rho}$. Let us define $\widehat{\rho}$ as follows:

$$\widehat{\rho} := \sup_{\theta:\mu_1(\theta)>2}\left[\frac{1-G_x\left(\frac{1+\overline{s}(\theta)}{2},\theta\right)}{\mu_1(\theta)-G_x\left(\frac{1+\overline{s}(\theta)}{2},\theta\right)}\right],$$

where $\overline{s}(\theta) \in (0,1)$ is the interior solution of the following fixed point condition:

$$s = G\left(\frac{1+s}{2},\theta\right).$$

It is readily verified that $\overline{s}(\theta)$ is well defined iff $\mu_1(\theta) = G_x(1,\theta) > 2$. Furthermore, from $G_x(\cdot,\theta) > 0$, we have:

$$G\left(\frac{1+s}{2},\theta\right) > G(s,\theta) \quad \forall s \in (0,1),$$

hence $s^*(\rho,\theta) < \overline{s}(\theta$,and, from $G_{xx}(\cdot,\theta) > 0$,

$$G_x\left(\frac{1+s^*(\rho,\theta)}{2},\theta\right) < G_x\left(\frac{1+\overline{s}(\theta)}{2},\theta\right).$$

Step 2: proof that $\widehat{\rho} > 0$. As the denominator in $\frac{1-G_x\left(\frac{1+\overline{s}(\theta)}{2},\theta\right)}{\mu_1(\theta)-G_x\left(\frac{1+\overline{s}(\theta)}{2},\theta\right)}$ is unambiguously positive whenever $\overline{s}(\theta)$ is well defined, it suffices to show that $1-G_x\left(\frac{1+\overline{s}(\theta)}{2},\theta\right) > 0$ for a non-degenerate domain of values of θ. This follows immediately from the following considerations:

$$\begin{array}{ccc}
\lim_{\theta\to\overline{\theta}} G(\cdot,\theta) = 0 & \text{and} & \lim_{\theta\to\overline{\theta}} G_x(\cdot,\theta) = 0 \\
\Downarrow & & \Downarrow \\
\lim_{\theta\to\overline{\theta}} \overline{s}(\theta) = 0 & \Longrightarrow & \lim_{\theta\to\overline{\theta}} G_x\left(\frac{1+\overline{s}(\theta)}{2},\theta\right) = 0.
\end{array}$$

Thus, $1-G_x\left(\frac{1+\overline{s}(\theta)}{2},\theta\right) > 0$ when θ is large enough, hence $\widehat{\rho} > 0$.

Step 3: existence of (θ_1,θ_2). We now prove that, for any $\rho < \widehat{\rho}$, there is a non-degenerate interval (θ_1, θ_2) such that

$$\theta_1 < \theta < \theta_2 \Longrightarrow y^*(\theta) = 0 \text{ and } 1 - s^*(\rho, \theta) > 0.$$

Let us fix some $\theta_0 \in \left(0, \overline{\theta}\right)$, such that

$$\rho < \frac{1 - G_x\left(\frac{1+\overline{s}(\theta_0)}{2}, \theta_0\right)}{\mu_1(\theta_0) - G_x\left(\frac{1+\overline{s}(\theta_0)}{2}, \theta_0\right)}.$$

Such θ_0 exists because $\frac{1-G_x\left(\frac{1+\overline{s}(\theta)}{2},\theta\right)}{\mu_1(\theta)-G_x\left(\frac{1+\overline{s}(\theta)}{2},\theta\right)}$ is a continuous function which takes on all values in $(0, \widehat{\rho})$, hence it also takes on all values in $(\rho, \widehat{\rho})$. Furthermore, because the inequality is strict, there is an open neighborhood (θ_1, θ_2) of θ_0, such that

$$\rho G_x'(1, \theta) + (1 - \rho) G_x'\left(\frac{1 + \overline{s}\,(\theta)}{2}, \theta\right) < 1$$

for all $\theta \in (\theta_1, \theta_2)$.

As shown above, $s^*\,(\rho, \theta) < \overline{s}\,(\theta)$ hence

$$\rho G_x'(1, \theta) + (1 - \rho) G_x'\left(\frac{1 + s^*\,(\rho, \theta)}{2}, \theta\right) < 1 \tag{32}$$

for all $\theta \in (\theta_1, \theta_2)$. This proves (ii).

(iii) The result follows immediately if we set $\theta \to \overline{\theta}$ in (15) – (16) and notice that, from Assumption 2, $\lim_{\theta \to \overline{\theta}} G(x, \theta) = 0$. This proves (iii) and completes the proof. □

Proof of Proposition 7.

(i) Under the exponential degree distribution, the fraction $s^*(\theta, \rho)$ of individuals who abstain from sharing a recommendation in the population in a polarized equilibrium (i.e., under $y = 0$) can be expressed in closed form as follows:

$$s^*(\theta, \rho) = \frac{2 - \theta - \sqrt{4\rho - 4\rho\theta + \theta^2}}{2\theta}.$$

Hence, the condition (32) for stability of polarized equilibrium becomes

$$\phi(\rho, \theta) < 1, \tag{33}$$

where $\phi(\rho, \theta)$ is defined by

$$\phi(\rho, \theta) := \rho \frac{1}{1 - \theta} + (1 - \rho) \frac{16\,(1 - \theta)}{\left(2 - \theta + \sqrt{4\rho - 4\rho\theta + \theta^2}\right)^2}. \tag{34}$$

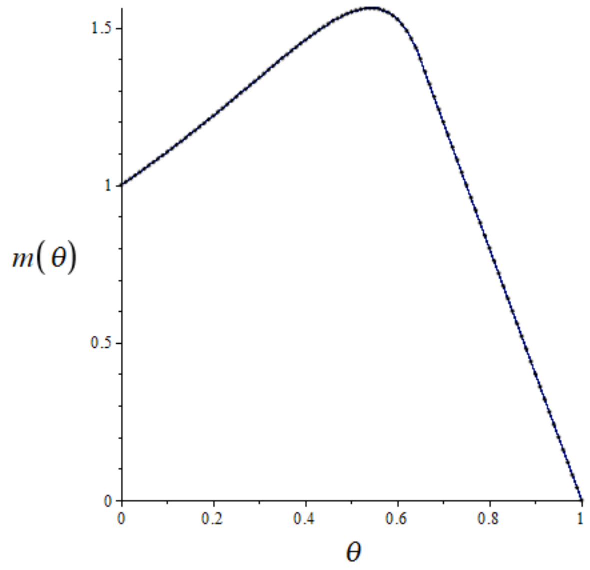


Figure 4: How $m(\theta) := \min_{\rho\in[0,1]} \phi(\rho,\theta)$ varies with θ.

We need the following Lemma.

Lemma 3. *The function $\phi(\rho,\theta)$ has the following properties:*

(a) $\theta \le 0.75 \Longrightarrow \phi(\rho,\theta) > 1$ *for all* $\rho \in (0,1)$*;*

(b) $0.75 < \theta < 1 \Longrightarrow \phi(0,\theta) < 1 < \phi(1,\theta) < 1$ *and* $\phi_\rho(\rho,\theta) > 0\ \forall\rho \in (0,1)$.

Proof of Lemma 3. (a) By plotting $m(\theta) = \min_{\rho\in[0,1]} \phi(\rho,\theta)$ as a function of θ over $(0,1)$, one finds that

$$\min_{\rho\in[0,1]} \phi(\rho,\theta) \gtreqless 1 \Longleftrightarrow \theta \lesseqgtr 0.75,$$

as one can see from Figure 4. This proves (a).

(b) Evaluate $\phi(\rho,\theta)$ at $\rho \in \{0,1\}$:

$$\phi(0,\theta) = 4\,(1-\theta) \lesseqgtr 1 \Longleftrightarrow \theta \gtreqless 0.75;$$

$$\phi(1,\theta) = \frac{1}{1-\theta} > 1.$$

This proves that $0.75 < \theta < 1 \Longrightarrow \phi(0,\theta) < 1 < \phi(1,\theta) < 1$. To prove that $\phi_\rho(\rho,\theta) > 0$ $\forall\rho \in (0,1)$, it suffices to show that

$$\phi_\rho(\rho,0.75) > 0 \qquad \text{and} \qquad \phi_{\rho\theta}(\rho,\theta) > 0.$$

The inequality $\phi_\rho(\rho,0.75) > 0$ means that $\phi(\rho,0.75)$ in increasing with respect to ρ, which one can establish by plotting $\phi(\rho,0.75)$ as a function of ρ over $[0,1]$, which is increasing in ρ (Figure 5).

As for the inequality $\phi_{\rho\theta}(\rho,\theta) > 0$, one can verify it by direct calculation. This proves (b) and completes the proof of Lemma 3. $\square$

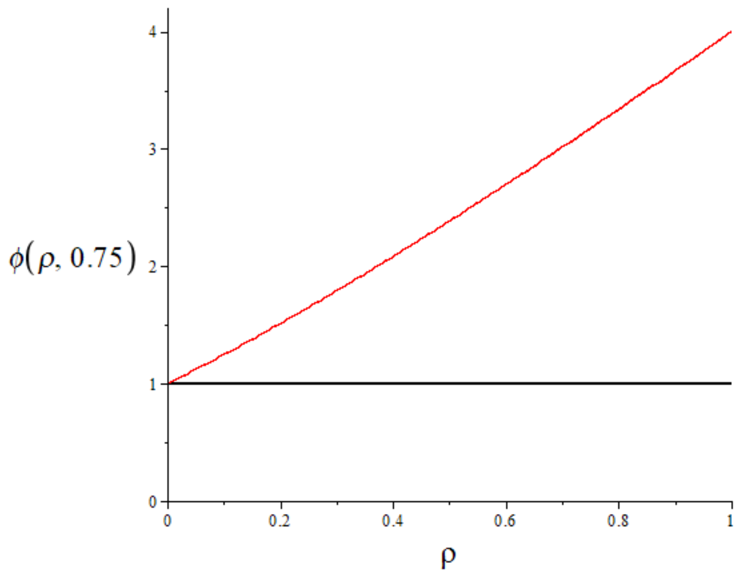


Figure 5: How $\phi(\rho, 0.75)$ varies with ρ.

We now proceed with the proof of Proposition 7(i). From Lemma 3(a), the stability condition (33) or the polarized equilibrium fails to hold when for $\theta \leq 0.75$. From Lemma 3(b), (33) has a unique interior solution $\rho_0^{\text{silent}}(\theta)$ with respect to ρ by the intermediate value theorem. This proves (i).

(ii) We will now show that an interior equilibrium (y^*, s^*) exists and is unique iff $\rho_0^{\text{silent}}(\theta) < \rho < \rho_1^{\text{silent}}(\theta)$. Observe first that, since the RHS of the dynamic equation (16) is decreasing in y_{t-1}, we have:

$$s_t \gtreqless s_{t-1} \Longleftrightarrow y_{t-1} \lesseqgtr \tilde{y}\left(s_{t-1}\right), \tag{35}$$

where

$$\tilde{y}(s) := 1 + s - \frac{2s\left(1-\rho+\rho\theta\right) - 2\theta s^2}{\left(1-\rho\right)\left(1-\theta\right) + \theta^2 s - \theta^2 s^2}. \tag{36}$$

Similarly, because the RHS of the dynamic equation (15) is increasing with respect to s_t, we have:

$$y_t \gtreqless y_{t-1} \Longleftrightarrow s_{t-1} \gtreqless \widetilde{s}(y_{t-1}), \tag{37}$$

where

$$\widetilde{s}(y) := \frac{2}{\theta}\left[1 - \sqrt{\left(\frac{1}{2}\theta y\right)^2 + \left(1-\rho\right)\left(1-\theta\right)\frac{1-\theta+\theta y}{\left(1-\theta+\theta y\right)-\rho}}\right] - 1. \tag{38}$$

The interior steady state, if it exists, must be an intersection point of two curves on the (s, y)-plane given by the equations:

$$y = \widetilde{y}(s), \qquad s = \widetilde{s}(y), \tag{39}$$

where the functions $\widetilde{y}(s)$ and $\widetilde{s}(y)$ are defined by, respectively, (36) and (38). From (35) and (37), it is clear that the two curves in (39) correspond to the loci of, respectively, $y_t = y_{t-1}$ and

$s_t = s_{t-1}$. Furthermore, one can show that the function $\widetilde{s}(y)$ is always concave, while the function $\widetilde{s}(y)$ is always decreasing and convex. Hence, the two curves have at most one interior intersection point. The necessary and sufficient conditions for the two curves to have an interior intersection point are as follows. First, the $\widetilde{s}(y)$ curve must be less steep at $(y, s) = (0, 1)$ than the $\widetilde{y}(s)$ curve:

$$\frac{1 - \frac{1}{2}\theta}{\frac{1}{2}\theta - (1-\theta)\frac{\rho}{1-\rho}} < \frac{1}{1-\theta}\left(1 + \theta + 2\theta\frac{\rho}{1-\rho}\right). \tag{40}$$

Second, the $\widetilde{s}(y)$ curve must intersect the s-axis closer to the origin than the $\widetilde{y}(s)$ curve

$$\widetilde{s}(0) < s^*(\theta, \rho) = \frac{2 - \theta - \sqrt{4\rho - 4\rho\theta + \theta^2}}{2\theta}, \tag{41}$$

where $s^*(\theta, \rho)$ is the polarized equilibrium (see proof of (i) above) which can be determined as the solution to $\widetilde{y}(s) = 0$.

One can show that (40) and (41) are equivalent, respectively, to $\rho < \rho_1^{\text{silent}}(\theta)$ and $\rho > \rho_0^{\text{silent}}(\theta)$. The stability of the interior equilibrium can be established qualitatively by means of the phase diagram which we provide on Figure 3c. The directions of the dynamics are determined by (37) and (35). This proves (ii).

(iii) The expression $\rho_0^{\text{silent}}(\theta) = \frac{2\theta-1}{4\theta-2\theta^2-1}$ for the exponential degree distribution can be obtained by plugging $p_1(\theta) = 1 - \theta$ and $\mu_1(\theta) = \frac{1}{1-\theta}$ into (30). Thus, (iii) follows immediately from Proposition 5(i).

(iv) The value $\widehat{\rho}^{\text{silence}} \approx 0.0685$ can be determined by finding numerically the solution to the equation

$$\min_{\theta \in [0,1]} \phi(\rho, \theta) = 1,$$

both sides of which are plotted in Figure 6.

From Figure 6, one can see that the stability condition (33) holds for a non-empty set of the network density levels iff $\rho < \widehat{\rho}^{\text{silence}}$. That set is given by $\{\theta : \phi(\rho, \theta) < 1\}$ and is an open interval $\left(\theta_2^{\text{silent}}, \theta_3^{\text{silent}}\right)$ because $\phi(\rho, \theta)$ has a unique minimizer w.r.t θ, hence it is quasi-convex with respect to θ. The rest of the proof follows from Propositions 5 and 6. This proves (iv) and completes the proof. $\square$

Proof of Proposition 8.

(i) The steady-state conditions of the two-dimensional dynamical system (19)-(20) are given by:

$$\ell = \left(\frac{1-\rho}{2} - \epsilon\right)(1 - G(1-\ell)) + \rho\frac{\ell}{\ell + r}G(\ell + r) + \left(\frac{1-\rho}{2} + \epsilon\right)G(\ell); \tag{42}$$

$$r = \left(\frac{1-\rho}{2} + \epsilon\right)(1 - G(1-r)) + \rho\frac{r}{\ell + r}G(\ell + r) + \left(\frac{1-\rho}{2} - \epsilon\right)G(r). \tag{43}$$

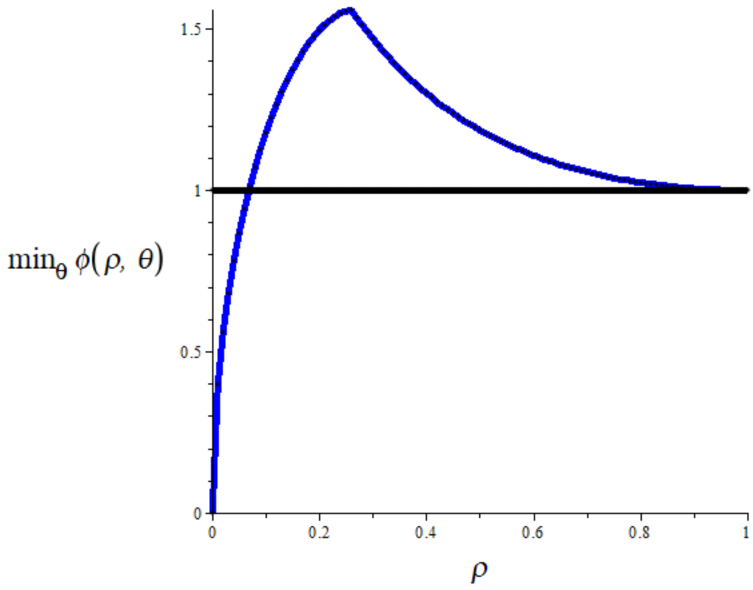


Figure 6: How $\min_\theta \phi(\rho, \theta)$ varies with ρ.

It is readily verified that the non-polarized equilibrium $(\ell^*, r^*) = (0, 0)$ satisfies (42)-(43). The Jacobian of the dynamical system (19)-(20) evaluated at $(\ell^*, r^*) = (0, 0)$ is a diagonal matrix given by

$$\begin{pmatrix} \left(\frac{1-\rho}{2} - \epsilon\right) G'(1) + \left(\frac{1+\rho}{2} + \epsilon\right) G'(0) & 0 \\ 0 & \left(\frac{1-\rho}{2} + \epsilon\right) G'(1) + \left(\frac{1+\rho}{2} - \epsilon\right) G'(0) \end{pmatrix}.$$

The eigenvalues of the Jacobian are as follows:

$$\lambda_{1,2}(\rho, \epsilon) = \left(\frac{1+\rho}{2} \pm \epsilon\right) p_1 + \left(\frac{1-\rho}{2} \mp \epsilon\right) \mu_1.$$

As we focus on the case of right-skewed preference distributions ($\epsilon > 0$), the condition for the non-polarized equilibrium to be stable is $\lambda_2(\rho, \epsilon) < 1$. Or, equivalently:

$$\rho > \rho_1(\epsilon) := \frac{\mu_1 + p_1 - 2}{\mu_1 - p_1} + 2\epsilon.$$

This proves (i).

(ii) The polarized equilibrium (ℓ^*, r^*) satisfies the following property:

$$\ell^* + r^* = 1.$$

Hence, the steady-state conditions (42) – (43) become:

$$\ell = \left(\frac{1-\rho}{2} - \epsilon\right)(1 - G(1-\ell)) + \rho\ell + \left(\frac{1-\rho}{2} + \epsilon\right) G(\ell);$$

$$r = \left(\frac{1-\rho}{2} + \epsilon\right)(1 - G(1-r)) + \rho r + \left(\frac{1-\rho}{2} - \epsilon\right) G(r).$$

If the first steady-state condition holds for some ℓ^*, then the second steady-state condition holds for $r^* = 1 - \ell^*$, and vice versa. The solutiuon $\ell^* \in (0,1)$ to the first steady-state condition exists if and only if

$$\left(\frac{1-\rho}{2} - \epsilon\right) G'(1) + \rho + \left(\frac{1-\rho}{2} + \epsilon\right) G'(0) > 1,$$

which, in turn, holds true if and only if

$$\rho < 1 - 2\epsilon \frac{\mu_1 - p_1}{\mu_1 + p_1 - 2}.$$

Because we assume $\epsilon > 0$ to be small, this condition is likely to be satisfied.

The Jacobian of the dynamical system (19)-(20) evaluated at the polarized equilibrium is given by:

$$\left(\begin{array}{cc} \left[\left(\frac{1-\rho}{2} + \epsilon\right) G'(\ell^*) + \left(\frac{1-\rho}{2} - \epsilon\right) G'(r^*) + \rho\right] + \rho\ell^* \left(G'(1) - 1\right) & \rho\ell^* \left(G'(1) - 1\right) \\ \rho r^* \left(G'(1) - 1\right) & \left[\left(\frac{1-\rho}{2} + \epsilon\right) G'(\ell^*) + \left(\frac{1-\rho}{2} - \epsilon\right) G'(r^*) + \rho\right] + \end{array} \right.$$

It is readily verified that the eigenvalues of this Jacobian are

$$\lambda_1(\rho,\epsilon) = \left(\frac{1-\rho}{2} + \epsilon\right) G'(\ell^*(\epsilon)) + \left(\frac{1-\rho}{2} - \epsilon\right) G'(r^*(\epsilon)) + \rho G'(1);$$

$$\lambda_2(\rho,\epsilon) = \left(\frac{1-\rho}{2} + \epsilon\right) G'(\ell^*(\epsilon)) + \left(\frac{1-\rho}{2} - \epsilon\right) G'(r^*(\epsilon)) + \rho < 1.$$

The steady state condition in this case is therefore $\lambda_1(\rho,\epsilon) < 1$. The threshold level $\rho_0(\epsilon)$ of ρ is the solution to

$\lambda_1(\rho,\epsilon) = 1$. When $\epsilon = 0$, the polarized equilibrium is perfectly symmetric,

$$(\ell^*(\epsilon), r^*(\epsilon)) = \left(\frac{1}{2}, \frac{1}{2}\right),$$

and one can readily verify that $\lambda_1(\rho,\epsilon)|_{\epsilon=0}$ is linear increasing with respect to ρ. By continuity, it must be that

$$\frac{\partial \lambda_1(\rho,\epsilon)}{\partial \rho} > 0$$

for some open neighborhood of $\epsilon = 0$. Hence, for ϵ not too large, there exists a unique solution $\rho_0(\epsilon)$ to $\lambda_1(\rho,\epsilon) = 1$ with respect to ρ.

It remains to verify that $\rho_0(\epsilon) < \rho_0(0)$ when p_2 is not too large. To see this, let us differentiate the principal eigenvalue $\lambda_1(\rho,\epsilon)$ of the Jacobian with respect to ϵ:

$$\frac{\partial \lambda_1(\rho,\epsilon)}{\partial \epsilon} = G'(\ell^*(\epsilon)) - G'(r^*(\epsilon)) + \left(\frac{1-\rho}{2} + \epsilon\right) G''(\ell^*(\epsilon)) \frac{d\ell^*(\epsilon)}{d\epsilon} + \left(\frac{1-\rho}{2} - \epsilon\right) G''(r^*(\epsilon)) \frac{dr^*(\epsilon)}{d\epsilon}.$$

Evaluating $\frac{\partial \lambda_1(\rho,\epsilon)}{\partial \epsilon}$ in the vicinity of perfect symmetry ($\epsilon = 0$), we get:

$$\left.\frac{\partial \lambda_1(\rho,\epsilon)}{\partial \epsilon}\right|_{\epsilon=0} = 0.$$

Hence, we need to evaluate the second derivative of $\lambda_1(\rho,\epsilon)$ with respect to ϵ:

$$\frac{\partial^2 \lambda_1(\rho,\epsilon)}{\partial \epsilon^2} = 2\left[G''\left(\ell^*(\epsilon)\right)\frac{d\ell^*(\epsilon)}{d\epsilon} - G''\left(r^*(\epsilon)\right)\frac{dr^*(\epsilon)}{d\epsilon}\right]$$

$$+\left(\frac{1-\rho}{2}+\epsilon\right)G'''\left(\ell^*(\epsilon)\right)\left[\frac{d\ell^*(\epsilon)}{d\epsilon}\right]^2 + \left(\frac{1-\rho}{2}-\epsilon\right)G'''\left(r^*(\epsilon)\right)\left[\frac{dr^*(\epsilon)}{d\epsilon}\right]^2$$

$$+\left(\frac{1-\rho}{2}+\epsilon\right)G''\left(\ell^*(\epsilon)\right)\frac{d^2\ell^*(\epsilon)}{d\epsilon^2} + \left(\frac{1-\rho}{2}-\epsilon\right)G''\left(r^*(\epsilon)\right)\frac{d^2 r^*(\epsilon)}{d\epsilon^2}.$$

Evaluating the second derivative in the vicinity of perfect symmetry ($\epsilon = 0$), we get after simplifications:

$$\left.\left[\frac{\partial^2 \lambda_1(\rho,\epsilon)}{d\epsilon^2}\right]\right|_{\epsilon=0} = 8G''\left(\frac{1}{2}\right)\left[\frac{1}{2}\frac{G'''\left(\frac{1}{2}\right)}{G''\left(\frac{1}{2}\right)}\frac{\frac{1}{2}-G\left(\frac{1}{2}\right)}{1-G'\left(\frac{1}{2}\right)} - 1\right]\frac{\frac{1}{2}-G\left(\frac{1}{2}\right)}{(1-\rho)\left[1-G'\left(\frac{1}{2}\right)\right]}.$$

Hence,

$$\left.\left[\frac{\partial^2 \lambda_1(\epsilon)}{\partial \epsilon^2}\right]\right|_{\epsilon=0} \gtreqless 0 \Longleftrightarrow \frac{\frac{1}{2}G'''\left(\frac{1}{2}\right)}{G''\left(\frac{1}{2}\right)}\frac{\frac{1}{2}-G\left(\frac{1}{2}\right)}{1-G'\left(\frac{1}{2}\right)} \gtreqless 1.$$

This inequality holds with ">" when p_2 is not too large. To see this, set $p_2 = 0$ first. We get:

$$\frac{1}{2}\frac{G'''\left(\frac{1}{2}\right)}{G''\left(\frac{1}{2}\right)}\frac{\frac{1}{2}-G\left(\frac{1}{2}\right)}{1-G'\left(\frac{1}{2}\right)} = \underbrace{\frac{\frac{3}{2}p_3 + \sum_{k=4}^{\infty}\left(\frac{k}{2}-1\right)(k-1)k\left(\frac{1}{2}\right)^{k-2}p_k}{3p_3 + \sum_{k=4}^{\infty}(k-1)k\left(\frac{1}{2}\right)^{k-2}p_k}}_{>1/2} \times \underbrace{\frac{3p_3 + 4\sum_{k=4}^{\infty}p_k\left[1-\left(\frac{1}{2}\right)^{k-1}\right]}{p_3 + 4\sum_{k=4}^{\infty}p_k\left[1-k\left(\frac{1}{2}\right)^{k-1}\right]}}_{>3} > 1.$$

By continuity, the same is true if p_2 is positive but not too large. In this case, $\lambda_1(\rho,\epsilon)$ increases with both ρ and ϵ in the vicinity of $\epsilon = 0$, and thus $\rho_0(\epsilon)$, which solves $\lambda_1(\rho,\epsilon)$, increases in response to a bit more asymmetry. This proves (ii) and completes the proof. □

References

Algan, Y., N. Dalvit, Q.-A. Do, A. Le Chapelain, and Y. Zenou (2025). Friendship networks and political opinions. Unpublished manuscript, Monash University.

Allcott, H., L. Braghieri, S. Eichmeyer, and M. Gentzkow (2020). The welfare effects of social media. *American Economic Review 110*(3), 629–676.

Allcott, H. and M. Gentzkow (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives 31*, 211–235.

Bail, C. A., L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences 115*(37), 9216–9221.

Barberá, P. (2015). How social media reduces mass political polarization. Evidence from Germany, Spain, and the US. Unpublished manuscript, New York University.

Barberá, P. (2020). Social media, echo chambers, and political polarization. In: N. Persily and J. Tucker*, Social Media and Democracy: The State of the Field, Prospects for Reform,* Cambridge University Press, Cambridge, UK, 34–55.

Barberá, P., J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science 26*(10), 1531–1542.

Bolletta, U. and P. Pin (2025). Dynamic opinion updating with endogenous networks. *European Economic Review,* forthcoming.

Boxell, L., M. Gentzkow, and J. M. Shapiro (2017). Greater internet use is not associated with faster growth in political polarization among us demographic groups. *Proceedings of the National Academy of Sciences 114*(40), 10612–10617.

Callander, S. and J. C. Carbajal (2022). Cause and effect in political polarization: A dynamic analysis. *Journal of Political Economy 130*(4), 825–880.

Campbell, A. (2013). Word-of-mouth communication and percolation in social networks. *American Economic Review 103*(6), 2466–2498.

Campbell, A., P. Ushchev, and Y. Zenou (2024). The network origins of entry. *Journal of Political Economy 132*(11), 3867–3916.

Della Lena, S., L. P. Merlino, and Y. Zenou (2023). Affective polarization, media outlets, and opinion dynamics. CEPR Discussion Paper No. 18508.

Dubois, E. and G. Blank (2018). The echo chamber is overstated: the moderating effect of political interest and diverse media. *Information, Communication & Society 21*(5), 729–745.

El-Bermawy, M. M. (2016). Your filter bubble is destroying democracy. *Wired, November 18.*

Guess, A. M., N. Malhotra, J. Pan, P. Barberá, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Dimmery, D. Freelon, M. Gentzkow, et al. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science 381*(6656), 398–404.

Hindman, M. (2009). *The Myth of Digital Democracy.* Princeton University Press.

Levy, G. and R. Razin (2019). Echo chambers and their effects on economic and political outcomes. *Annual Review of Economics 11*(1), 303–328.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review 111*(3), 831–870.

Mutz, D. C. (2006). *Hearing the Other Side: Deliberative versus Participatory Democracy.* Cambridge University Press.

Newman, M. E. (2010). *Networks. An Introduction.* Oxford: Oxford University Press.

Newman, M. E., S. H. Strogatz, and D. J. Watts (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E 64*(2), 026118.

Newman, M. E., D. J. Watts, and S. H. Strogatz (2002). Random graph models of social networks. *Proceedings of the National Academy of Sciences 99*, 2566–2572.

Pariser, E. (2011). *The Filter Bubble: What the Internet is Hiding from you.* penguin UK.

Ross Arguedas, A., C. Robertson, R. Fletcher, and R. Nielsen (2022). Echo chambers, filter bubbles, and polarisation: A literature review. Unpublished manuscript.

Sunstein, C. R. (2018). *Republic: Divided Democracy in the Age of Social Media.* Princeton: Princeton University Press.

Sustein, C. R. (2008). *Republic. com 2.0.* Princeton: Princeton University Press.

Terren, L. and R. Borge (2021). Echo chambers on social media: A systematic review of the literature. Unpublished manuscript.

Vega-Redondo, F. (2007). *Complex Social Networks.* Cambridge: Cambridge University Press.

Online Appendix: Additional results

A Verifying $\rho_0 \leq \rho_1$

It is readily verified that $\mu_1 + p_1 = \mathbb{E}\left[\max\left\{\widetilde{k}, 2\right\}\right] \geq 2$, with ">" if $p_k > 0$ for some $k \geq 3$. Hence $\mu_1 - 1 \geq 1 - p_1$. Furthermore, from Lemma B1, $\mu_1 + p_1 - 2 \geq 2\left(1 - \psi\right)$. Hence, the following inequality holds:

$$\left(\mu_1 + p_1 - 2\right)\left(\mu_1 - 1\right) \geq 2\left(1 - \psi\right)\left(1 - p_1\right).$$

Or, after equivalent transformations and using (6) – (7):

$$\rho_0 = \frac{1 - \psi}{\mu_1 - \psi} \leq \frac{\mu_1 + p_1 - 2}{\mu_1 - p_1} = \rho_1.$$

with "<" if $p_k > 0$ for some $k \geq 3$. This completes the proof. □

B A useful Lemma

Lemma B1. *For any degree distribution $\{p_k\}$ over $\mathbb{N}$ satisfying $\mu_1 < \infty$, the following inequality holds*

$$\mu_1 + p_1 + 2\psi \geq 4, \tag{B.1}$$

with ">" if $p_k > 0$ for some $k \geq 3$, i.e., if the fraction of individuals having at least three friends is positive.

Proof of Lemma B1. We have:

$$\mu_1 + p_1 = \mathbb{E}\left[\max\left\{\widetilde{k}, 2\right\}\right];$$

$$\psi = \mathbb{E}\left[\widetilde{k}\left(\frac{1}{2}\right)^{\widetilde{k}-1}\right].$$

Hence, (B.1) will follow if we show that the inequality

$$\max\left\{k, 2\right\} + k\left(\frac{1}{2}\right)^{k-2} \geq 4 \tag{B.2}$$

holds holds for all $k \in \mathbb{N}$, with ">" if $p_k > 0$ for some $k \geq 3$. For $k \in \{1, 2\}$, (B.2) holds with "=". For $k = 3$, the LHS of (B.2) equals $\frac{9}{2} > 4$, hence (B.2) holds with ">". Finally, for $k \geq 4$, we have:

$$\max\{k,2\} + k\left(\frac{1}{2}\right)^{k-2} = k\left(1+\left(\frac{1}{2}\right)^{k-2}\right) > k \geq 4.$$

This completes the proof. □

C Multiple switches between partial polarization and no polarization under an FOSD shock

In this section, we show how to develop examples of the following sort: there are three degree distributions, $\{p_k^0\}$, $\{p_k^1\}$, and $\{p_k^2\}$, satisfying

$$\{p_k^0\} \prec_{\text{FOSD}} \{p_k^1\} \prec_{\text{FOSD}} \{p_k^2\},$$

and such that there is partial polarization under $\{p_k^0\}$, no polarization under $\{p_k^1\}$, and partial polarization again under $\{p_k^2\}$. To see this, let us arbitrarily choose $\{p_k^0\}$, with the only restrictions that $p_1^0 > 0$, and $p_k^0 > 0$ for some $k \geq 3$. Let ρ be slightly below ρ_1 under $\{p_k^0\}$. Let us choose $\{p_k^1\}$ as follows:

$$p_k^1 = \begin{cases} p_k^0 - \epsilon, & k=1; \\ p_k^0 + \epsilon, & k=2; \\ p_k^0, & k \geq 3. \end{cases}$$

It is readily verified that such a shock leads to a reduction of $\rho_1 = \frac{\mu_1+p_1-2}{\mu_1+p_1-2p_1}$, since $\mu_1 + p_1$ does not change while p_1 decreases. Next, let us choose $\{p_k^2\}$ by transferring some mass of $\{p_k^1\}$ rightwards, but so that $\mathbb{P}\left[\widetilde{k}=1\right]$ remains unchanged, i.e., $p_1^1 = p_1^2$. Then, one can see that ρ_1 increases, since it is an increasing function of μ_1. In particular, we can construct these FOSD-respecting shocks so that:

$$\rho_1\left(\{p_k^2\}\right) = \rho_1\left(\{p_k^0\}\right) < \rho < \rho_1\left(\{p_k^1\}\right).$$

Thus, two consecutive FOSD-respecting changes in the degree distribution can lead, first, to switching from partial polarization to no polarization, and then back to partial polarization. This is another illustration of social dynamics being highly non-monotone with respect to the network density.

D Full silence is never stable

The full silence equilibrium is the one where no one is active:

$$(y^*, s^*) = (0, 1) \, .$$

In this steady state, the Jacobian (29) takes the form

$$\begin{pmatrix} G'(1) & 0 \\ -\frac{1-\rho}{2} G'(1) & \left(\rho + \frac{1-\rho}{2}\right) G'(1) \end{pmatrix},$$

hence the eigenvalues of the Jacobian are:

$$\lambda_1 = G'(1), \qquad \lambda_2 = \left(\rho + \frac{1-\rho}{2}\right) G'(1) \, .$$

As $\lambda_1 = G'(1) > 1$, the full silence equilibrium is never a stable outcome. This completes the proof. $\square$