ROCKWOOL Foundation Berlin
Institute for the Economy and the Future of Work (RFBerlin)

DISCUSSION PAPER SERIES 82/25

The Polarization Paradox: Why More Connections Can Divide Us

Arthur Campbell, C. Matthew Leister, Philip Ushchev, Yves Zenou

www.rfberlin.com

OCTOBER 2025

The Polarization Paradox: Why More Connections Can Divide Us

Authors

Arthur Campbell, C. Matthew Leister, Philip Ushchev, Yves Zenou

Reference

JEL Codes: D83, D85, L82

Keywords: Social networks, network density, content filtering, polarization

Recommended Citation: Arthur Campbell, C. Matthew Leister, Philip Ushchev, Yves Zenou (2025): The Polarization Paradox: Why More Connections Can Divide Us. RFBerlin Discussion Paper No. 82/25

Imprint

RFBerlin
ROCKWOOL Foundation Berlin –
Institute for the Economy
and the Future of Work

Gormannstrasse 22, 10119 Berlin
Tel: +49 (0) 151 143 444 67
E-mail: info@rfberlin.com
Web: www.rfberlin.com

The Polarization Paradox: Why More Connections Can Divide Us*

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

October 9, 2025

Abstract

We develop a simple model of *content filtering*—the tendency of individuals to selectively forward information that aligns with their ideological preference—to study how network structure shapes the distribution of political content. In our framework, individuals and content are horizontally differentiated into three types (left, middle, right). We show that content filtering can amplify the middle or the extremes and may result in only centrist content (full moderation) or only extreme content (full polarization). The outcome depends on the interaction between two forces: a *preference advantage* from the relative prevalence of types in the population, and a *pairwise comparison advantage* that systematically favors centrist content. Network density plays a critical role. Sparse networks robustly yield moderation, even when extreme types dominate the population, while dense networks replicate the population's type distribution. Intermediate densities generate non-monotonic comparative statics, including sharp transitions between moderation and polarization. These findings complement existing empirical results that emphasize the types of connections individuals have on social media by highlighting how the *number* of connections, holding their composition fixed, may fundamentally shape the information environment in ways that foster/mitigate populism and polarization.

JEL classification numbers: D83, D85, L82.

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

*We thank Jean-Marie Baland, Laurent Bouton, Ben Golub, Federico Masera, David McAdams, Moritz Meyer-ter-Vehn, Torsten Persson, Jean-Philippe Platteau, Guido Tabellini and the participants of the Monash Network Conference (Melbourne, 12-13 September 2025), the 29th Annual ISNIE / SIOE Conference (Sydney, 24-26 August 2025), the BSE Summer Forum Workshop on Networks (Barcelona, 16-17 June 2025), the Asian-Pacific Industrial Organization 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 Georgetown University, the University of New South Wales, 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 and social media in shaping political beliefs and behaviors through how they affect the information available to individuals. This literature has focused on how social media shapes the types of relationships that are formed and how this influences the type(s) of information available to users. One prevalent view links the lower costs of interacting online through social media to the formation of "echo chambers"—environments in which individuals are primarily exposed to ideologically aligned individuals and content.[1] A more recent wave of empirical studies questions whether the echo chamber mechanism is empirically validated and rather finds that social media can, under certain conditions, expand users' exposure to more diverse information and opinions beyond immediate social circles via weak-ties.[2] In both perspectives, the central premise is that individuals are more likely to share information consistent with their ideological preferences. The key distinction between the two lies in how increased social media use shapes the composition of one's social environment—and, consequently, the nature of the information encountered. The echo-chamber view posits that social media reinforces homophily by encouraging interactions among like-minded individuals, whereas the weak-ties view argues the opposite, suggesting that social media expands exposure to diverse viewpoints. The coexistence of these seemingly opposing mechanisms—both supported by empirical evidence—underscores the need for a more nuanced theoretical framework linking network structure to information exposure. We propose an alternative mechanism to explain how social media shapes polarization, focusing on individuals' selective tendencies in deciding which information to share. We refer to this process as content filtering.[3] The key insight of our model is that changes in network density can either amplify or dampen polarization, even when the overall composition of available information remains constant—a dimension largely overlooked in the existing empirical literature.

To study this, we develop a simple model of content filtering on social media and characterize the resulting steady-state distribution of shared content in a setting with three types of information: Left, Middle, and Right. Despite its simplicity, the model generates a rich set of outcomes and offers an alternative mechanism to reconcile the aforementioned empirical findings. In particular, more intense use of social media (captured in our model by increased network density) may moderate or polarize the steady-state distribution of information available to the population. This non-monotonicity may arise in our model because of the counteracting effects of a pairwise-comparison advantage of the middle content and a preference advantage of the left/right extreme

[1] See, e.g., Mutz (2006), Hindman (2009), Pariser (2011), Sustein (2008); Sunstein (2018), El-Bermawy (2016), Allcott and Gentzkow (2017), Allcott et al. (2020), and Mueller (2025).

[2] See, e.g., Bakshy et al. (2015), Barberá (2015); Barberá et al. (2015); Boxell et al. (2017); Barberá (2020); Dubois and Blank (2018); Algan et al. (2025).

[3] In an experiment, Messing and Westwood (2014) show that Republicans select Fox News at a substantially higher rate than other sources, while Democrats are more likely to select MSNBC.

content. Changes in network density affects both the overall and relative strength of each creating the non-monotonicity. Our mechanism focuses on the number of sources of information on social media as distinct from the "echo-chamber" and "weak-ties' views which focus on the types of sources/connections to explain how social media use affects the availability of information.

Our framework features a horizontally differentiated population of agents and information content which are one of three ideological types—left, middle, and right—located on a line.[4] Our model of content filtering is dynamic. In each period individuals observe the recommendations/ forwarded content of their friends and choose the content that is closest to their own type to subsequently recommend/forward in the subsequent period. We characterize the unique stable steady state of this process as a function of the distributions of types and friendships in the population.[5]

Our first set of results establishes the tendency for content filtering to amplify the prevalence of either the extremes (polarization) or the middle content (moderation) relative to the distribution of those types amongst the population.[6] The strength of this amplification may be such that the only type of content is the middle (full moderation) or the extremes (full polarization). Whether the middle or extremes are amplified depends on the balance of two sources of advantage in our model. One source of advantage (preference advantage) arises from a greater proportion of people in the population of that type. When the population has a higher fraction of a given type, the steady state will (weakly) reflect this. Moreover, when a type is more prevalent than another type this is one source of advantage that may result in the steady state amplifying that type of content. This first source of advantage may be reinforced or counteracted by a second source of advantage (pairwise comparison advantage) that always favors the middle content. This arises because the middle type of content is more likely to be favored over the left or right type in pairwise comparisons by randomly chosen individuals in the population.

Our second set of results demonstrates how the density of connections between people affects the steady state. Strikingly, for any distribution of preferences in the population the stable steady state is full moderation as the network becomes sparse. This occurs in networks where the preference advantage is arbitrarily high (i.e. as the proportion of extreme types goes to 1) indicating that the sparsity of the network allows the pairwise comparison advantage to dominate the preference advantage. In the other extreme, we find that dense networks weaken both sources of advantage and the steady-state converges to the proportion of each type in the population. In between, non-trivial comparative statics emerge when the preference advantage and pairwise comparison advantage counteract one another. Specifically, we find that the transition may (i) be non-monotonic transitioning through partially and potentially fully polarized steady states; and

[4]This is a 3 type version of Hotelling line used in models of spatial competition in political economy and industrial organization

[5]Our model of information diffusion draws on the theory of diffusion on random graphs (Newman et al., 2001, 2002) and build on its recent application in models of demand in imperfectly competitive markets in industrial organization (Campbell, 2013, 2015, 2019) and (Campbell et al., 2024)

[6]We use the term amplification of a type(s) to refer to steady states where there is a greater fraction of content of that type than there are types in the population.

(ii) exhibit a rapid transition between a fully moderate steady state and fully polarized steady state when there are few centrist/middle types and the network is relatively sparse.

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

As stated above, there are different views on how social networks affect polarization. Some argue that the changing environment—in which exposure to news is increasingly mediated through online social networks—has fostered the creation of "echo chambers," where individuals are exposed only to like-minded information and shielded from attitude-challenging content. Others take the opposite view, emphasizing that cross-cutting interactions are more frequent than commonly believed, so that social networking sites increase exposure to information shared by weak ties, including ideologically diverse news (Bakshy et al., 2015; Barberá et al., 2015). The empirical evidence, however, paints a more nuanced picture (Barberá et al., 2015; Arora et al., 2022). Deactivation experiments show that turning off Facebook or Instagram for several weeks slightly reduces affective polarization and political engagement (Allcott et al., 2020). Similarly, feed and intervention experiments—such as reducing like-minded content on Facebook during the 2020 U.S. election—altered information diets but produced only limited effects on polarization (Bakshy et al., 2015). At the macro level, observational evidence suggests that increasing internet and social-media penetration across U.S. demographic groups is not associated with a faster rise in polarization, pointing instead to weak or mixed links (Nyhan et al., 2023). Complementing these findings, Gentzkow and Shapiro (2011) assess the extent to which online news consumption is ideologically segregated and compare it with segregation in both traditional media and face-to-face interactions. They find that a substantial share of consumers access news from multiple outlets; for instance, visitors to extreme conservative sites such as rushlimbaugh.com and glennbeck.com are more likely than typical online readers to also visit nytimes.com, while visitors to extreme liberal sites such as thinkprogress.org and moveon.org are more likely than typical readers to visit foxnews.com. Similar patterns are documented by Messing and Westwood (2014).

Our contribution. We propose a new mechanism based on the intensity of content filtering to explain how the information available to individuals can evolve in ways that either foster or mitigate polarization. Proposition 1 shows that both full-polarization and full-moderation equilibria may arise, depending on the underlying parameter values. Furthermore, we demonstrate

that greater social media use—represented in the model by a denser network—can give rise to a non-monotonic evolution of polarization: beginning with full moderation, progressing to partial moderation, then shifting to full polarization, and ultimately stabilizing in a partial-polarization equilibrium (Proposition 3 and Figure 2).

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 evolution of the endogenous network. Although they show how homophily and clustering can endogenously produce polarization, we consider an exogenous network and vary its degree distribution parametrically. This allows us to isolate how the structural properties of the network, potentially driven by increased reliance on social media, may affect the equilibrium distribution of ideological information.

Our contribution. Our paper is the first to introduce a model of content filtering on social media. We show how changes in the network that describe the interactions of the population through social media can result in a distribution of information that can foster or inhibit polarized views. In focusing on the information available to the population as a whole, it offers a complementary mechanism to those considered in the previous literature.

The rest of the paper is organized as follows. Section 3 introduces our baseline model and some notations. Section 4.1 characterizes the steady-state equilibria while Section 4.2 presents the comparative statics with respect to the degree distribution. Sections 4.2.1, 4.3, and 4.4 establish 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 on social media which we use to study the impact of network structure and preference distribution on the information available to the population. We characterize the conditions under which content filtering amplifies moderate versus extreme content and vice-versa.

3.1 Notation, definitions, and preferences

Consider a model of ideological content 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 are connected to k friends from the previous period, where k is drawn from a degree distribution $\{p_k\}$ with expectation $\mu = \mathbb{E}[k]$ 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.[7] We see that under these preferences type M content has a relative advantage in pairwise

[7] 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.

comparisons against either of the extremes.

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 of recommended content. In each period t, every individual of any type connects to 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.[8] 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 $t-1$.[9] 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

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

[8] 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.

[9] In our model, we assume that a single recommendation from a same-type individual is sufficient for someone to recommend the same content. A natural extension would be to increase the threshold from one to two—that is, requiring at least two friends to recommend the same type of content before an individual follows suit. This modification would introduce the possibility of multiple stable steady states, thereby necessitating a different analytical approach. For instance, the analysis may need to shift toward comparing the relative sizes of the basins of attraction associated with different equilibria, a consideration that the current framework does not require.

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

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

4 Analysis

Our analysis focuses on stable steady-states $y^* \in [0,1]$ of the dynamical system (2). A state y^* is a steady state 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.

Definition 1. *A steady state y^* of the dynamical system (2) is stable if and only if any trajectory $\{y_t\}$ of the dynamical system (2) whose initial state is sufficiently close to y^*, asymptotically converges to y^*, that is, $\lim_{t\to\infty} y_t = y^*$. The stability criterion is given by:*

$$|f_y(y^*, \rho)| < 1,$$

where $f_y(y^*, \rho)$ is the partial derivative of $f(y, \rho)$ with respect to y evaluated at $y = y^*$.

Definition 1 corresponds to the standard notion of local stability in discrete-time dynamical systems. As we will show below, the dynamical system (2) admits a unique stable steady state, which is in fact globally stable. Thus, the "sufficiently close" condition in the definition can be omitted.

4.1 Steady-state characterization

Define the following two threshold values:[10]

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

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

where

$$\psi := G'\left(\frac{1}{2}\right). \tag{8}$$

The following proposition characterizes the stable steady-state fraction of content y^* that is forwarded by the population depending on the proportion of middle types ρ relative to extreme types $1-\rho$. It separates cases into steady-states where either the middle or extremes are amplified

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

relative to their composition of the population. We described steady-states where the middle is (extremes are) overrepresented $y^* > \rho$ ($y^* < \rho$) as full or partial moderation (full or partial polarization) and the threshold case where the steady-state reflects the population ($y^* = \rho$) as balanced.

Proposition 1. *There exists a $\tilde{\rho} < \frac{1}{3}$ where $y^*(\tilde{\rho}) = \tilde{\rho}$ such that the unique stable steady-state defined by (5) is characterized as follows:*

*(i) If $\rho \leq \rho_0$, there is **full polarization** of content, that is, $y^* = 0$.*

*(ii) If $\rho_0 < \rho < \tilde{\rho}$, there is **partial polarization** of content, that is, $0 < y^* < \rho$.*

*(iii) If $\rho = \tilde{\rho}$ content is **balanced**, that is, $y^*(\rho) = \rho$.*

*(iv) If $\tilde{\rho} < \rho < \rho_1$ there is **partial moderation** of content, that is, $y^* > \rho$.*

*(v) If $\rho \geq \rho_1$, there is **full moderation** of content, that is, $y^* = 1$.*

Moreover, y^* is strictly increasing in ρ for $\rho_0 < \rho < \rho_1$, i.e., $\frac{\partial y^*}{\partial \rho} := y^*_\rho > 0$.

We find that content filtering via forwarding behavior will tend to amplify one or other of the middle (moderation $y^* > \rho$) or the extreme (polarization $y^* < \rho$) types of content. Moreover, the strength of this amplification may result in either the middle/extremes fully crowding out the other, resulting in either full moderation ($y^* = 1$) or full polarization ($y^* = 0$) in populations that have a sufficiently high fraction of middle or extreme types respectively.

There are two sources of advantage at work in our model of content filtering that determine whether the middle type of content is amplified (moderation) or the extreme type of content is amplified (polarization). The first, *preference advantage*, is that individuals of each type recommend content that is similar to their own type when they observe it. Hence, when there is a greater fraction of individuals of that type in the population that content it is more likely to be recommended. The second, *pairwise comparison advantage*, arises when an individual does not observe content aligned with its own type. In that case the middle type of content has an advantage in pairwise comparisons against either of the extremes.[11] The combination of these two forces determines whether content filtering results in moderation or polarization. These may work in the same direction to favor the middle type ($\rho > \frac{1}{3}$) or in opposite directions ($\rho < \frac{1}{3}$). Indeed when the population contains an even fraction of all types $\rho = \frac{1}{3}$, there is no preference advantage across the types, so the pairwise advantage force leads the middle content to be amplified. Moreover, the threshold composition of the population where content filtering is balanced must be at a point where the two forces are working in opposite directions to offset each other, hence it is at a point where there are relatively more extreme types ($\tilde{\rho} < \frac{1}{3}$).

[11] Either extreme type of individual will choose to recommend the middle type of content over the alternative extreme type of content. However a middle type individual is equally likely to recommend either extreme type of content.

Figure 1 illustrates how the distribution of preferences maps into the distribution of behavior, where behavior reflects content recommendation patterns.[12] The blue bars in Figure 1 represent preference distributions, while the red bars correspond to content distributions. 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.

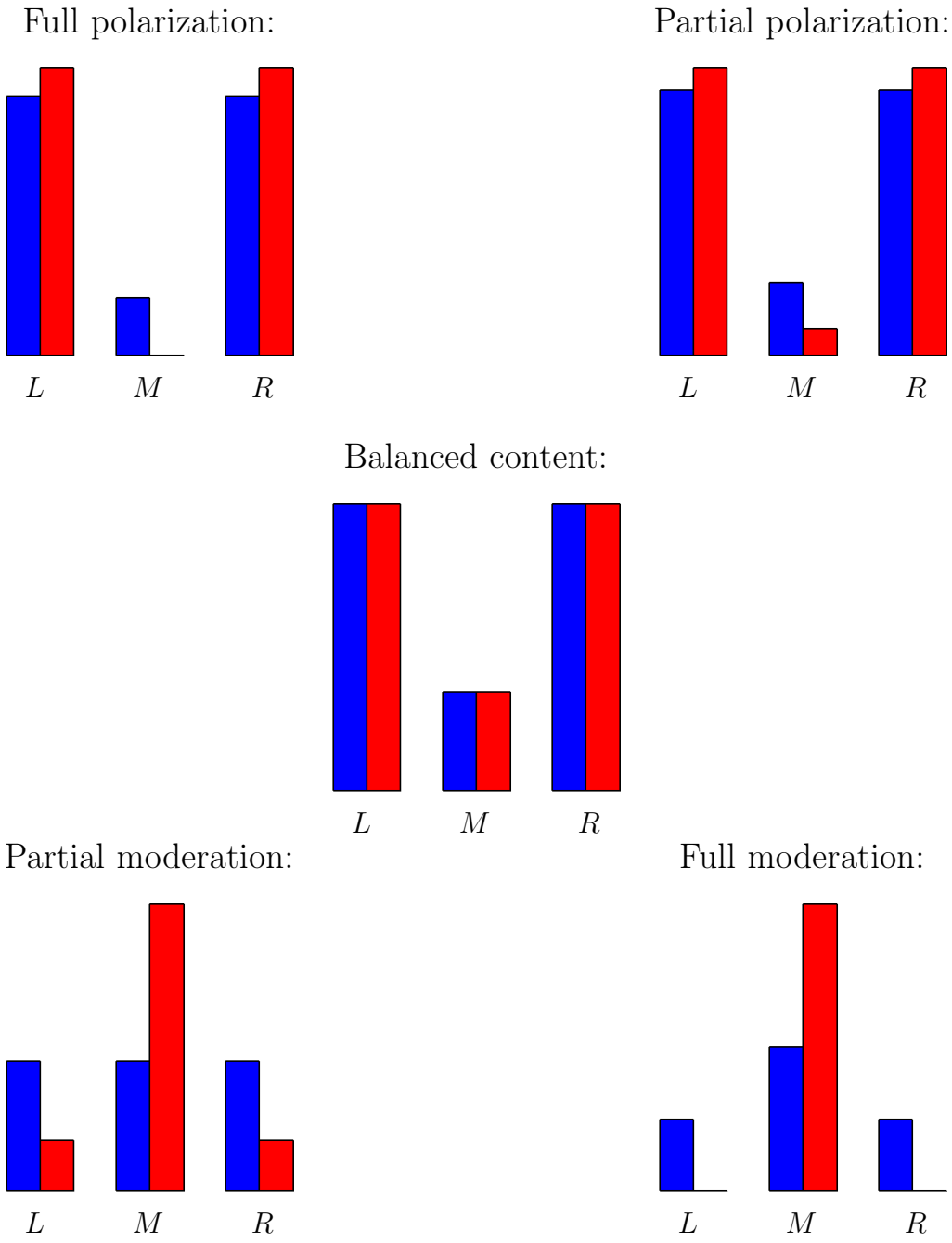


Figure 1: Preference distribution (blue bars) vs content distribution (red bars).

[12]Figure 1 is based on simulations using an exponential degree distribution (see equation (14) below) with density parameter $\theta = 0.6$. The values of the preference distribution parameter ρ corresponding to the five cases are: $\rho = 0.1$ for full polarization; $\rho = 0.12$ for partial polarization; $\rho = \widetilde{\rho} \approx 0.147$ for balanced content; $\rho = 1/3$ for partial moderation; and $\rho = 1/2$ for full moderation.

4.2 Comparative statics with respect to the degree distribution

Our broader goal is to understand how structural features of the social network affect whether content filtering results in moderation or polarization. In particular,

- we analyze how increased network density affects the steady-state in general networks before deriving sharper results by focusing on the exponential distribution;
- we investigate how a mean-preserving spread in the degree distribution affects polarization, comparing the outcomes in more regular versus more dispersed (irregular) networks.

4.2.1 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 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 $\theta' > \theta$.*

We refer to θ as the *network density parameter* whereby 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. As the network becomes infinitely dense, $\theta \to \overline{\theta}$, everyone's number of connections diverges.

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.

Proposition 2 characterizes the conditions on network density that result in full moderation (part(i)) and full polarization (part(ii)). Finally, in part (iii), it establishes that content becomes balanced as the network becomes dense.

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

(i) For each distribution of preferences $\rho \in (0,1)$, there exists a threshold $\widehat{\theta}(\rho)$ such that in sufficiently sparse networks there is full moderation:

$$\theta < \widehat{\theta}(\rho) \quad \Longleftrightarrow \quad y^*(\theta) = 1 \quad \textit{(full moderation).}$$

Moreover, as preferences become extreme, the threshold approaches a sparse network, i.e., $\lim_{\rho\to 0} \widehat{\theta}(\rho) = 0$.

(ii) There is a threshold distribution of preferences $\widehat{\rho} < \frac{1}{3}$ such that, for all more extreme preferences $\rho < \widehat{\rho}$, there exists a non-degenerate interval of network densities $(\theta_1(\rho), \theta_2(\rho))$ with $0 < \theta_1(\rho) < \theta_2(\rho) < \overline{\theta}$ such that there is full polarization:

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

Moreover, as preferences become extreme, the lower bound for full polarization approaches a sparse network, i.e., $\lim_{\rho\to 0} \theta_1(\rho) = 0$, and the upper bound approaches a dense network, i.e., $\lim_{\rho\to 0} \theta_2(\rho) = \overline{\theta}$.

(iii) As the network becomes infinitely dense, the stable distribution of content converges to the distribution of preferences in the population:

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

Proposition 2 reveals the rich comparative statics that emerge in our model. These comparative statics arise because network density affects both the absolute and relative strength of the *preference advantage* and *pairwise comparison advantage*. When the distribution of preferences is sufficiently extreme, these sources of advantage counteract one another. In sparse networks ($\theta < \widehat{\theta}(\rho)$), the dominant force in our model is the pairwise-comparison advantage of the middle resulting in the full moderation outcome (part (i)). Remarkably, this is true for all distributions of preferences ($\rho \in (0,1)$) indicating that network sparsity is sufficient to overcome any amount of preference advantage. As the network becomes denser, the *preference advantage* becomes stronger relative to the *pairwise-comparison advantage* resulting in a transition of the steady state from full moderation to full polarization over the range of densities $\theta \in \left(\widehat{\theta}(\rho), \theta_1(\rho)\right)$ (through partial moderation, balanced, and partial polarization steady states). For a range of densities $\theta \in (\widehat{\theta}, \theta_1)$, the preference advantage is sufficiently strong in both an absolute and relative sense to sustain the full polarization outcome (part (ii)). Finally, as the network becomes sufficiently dense both sources of advantage become weaker in an absolute sense and the steady state converges to the balanced steady-state $\lim_{\theta \to \bar{\theta}} y^*(\theta) = \rho$ (part(iii)).

When the distribution of preferences in the population features a larger proportion of the middle type ($\rho > \frac{1}{3}$), both sources of advantage favor moderation. In sufficiently sparse networks, this leads to full moderation. However, as the network becomes denser, both advantages are weakened, resulting in only partial moderation, with the steady state approaching a balanced configuration.

When the preference advantage favors the extremes but is not strong enough to induce full polarization ($\widehat{\rho} < \rho < \frac{1}{3}$), the transition from full moderation in sparse networks to a balanced outcome in dense networks may be non-monotonic, passing through steady states of partial polarization. While we cannot provide general results for arbitrary network structures, the next section offers a sharper characterization of the comparative statics under the exponential distribution.

Finally, as preferences become increasingly extreme ($\rho \to 0$), both the right and left thresholds for full moderation and full polarization — $\widehat{\theta}$ and θ_1, respectively — converge to zero: $\lim_{\rho \to 0} \widehat{\theta}(\rho) = \lim_{\rho \to 0} \theta_1(\rho) = 0$. This implies that the transition between full moderation and full polarization occurs over an increasingly narrow range of network densities, highlighting that in sparse networks with extreme preferences, the steady state can be highly sensitive to small changes in density.

Collectively, these results demonstrate that increases in network density can have non-monotonic and sometimes dramatic effects on the information environment.[13] In this way, our model reconciles the conflicting empirical evidence discussed in the introduction: the *echo-chamber view*, which suggests that social media drives individuals toward more homogeneous interactions, and the

[13] In the Online Appendix C, we provide a further illustration of the nonmonotone nature of the steady state with respect to network density. 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 moderation under $\{p_k^0\}$, full moderation under $\{p_k^1\}$, and partial moderation again under $\{p_k^2\}$.

weak-ties view, which argues that it instead exposes them to more diverse perspectives.

4.3 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 exists four thresholds $0 < \widehat{\theta}(\rho) < \widetilde{\theta}(\rho) < \theta_1(\rho) < \theta_2(\rho) < 1$ such that the stable steady-state $y^*(\theta,\rho)$ satisfies:

$$\begin{aligned}
\theta \le \widehat{\theta}(\rho) &\implies y^*(\theta,\rho) = 1 \quad \textit{(full moderation)},\\
\widehat{\theta}(\rho) < \theta < \widetilde{\theta}(\rho) &\implies \rho < y^*(\theta,\rho) < 1 \quad \textit{(partial moderation)},\\
\theta = \widetilde{\theta}(\rho) &\implies y^*(\theta,\rho) = \rho \quad \textit{(balanced)},\\
\widetilde{\theta}(\rho) < \theta < \theta_1(\rho) &\implies 0 < y^*(\theta,\rho) < \rho \quad \textit{(partial polarization)},\\
\theta_1(\rho) \le \theta \le \theta_2(\rho) &\implies y^*(\theta,\rho) = 0 \quad \textit{(full polarization)},\\
\theta > \theta_2(\rho) &\implies 0 < y^*(\theta,\rho) < \rho \quad \textit{(partial polarization)},\\
\theta \to 1 &\implies y^*(\theta,\rho) \to \rho \quad \textit{(balanced)}.
\end{aligned}$$

(ii) If $\frac{1}{9} \le \rho < \sqrt{5}-2 \approx 0.236$, then there exists thresholds $\widehat{\theta}(\rho)$ and $\widetilde{\theta}(\rho)$ such that the stable steady-state $y^*(\theta,\rho)$ satisfies:

$$\begin{aligned}
\theta \le \widehat{\theta}(\rho) &\implies y^*(\theta,\rho) = 1 \quad \textit{(full moderation)},\\
\widehat{\theta}(\rho) < \theta < \widetilde{\theta}(\rho) &\implies \rho < y^*(\theta,\rho) < 1 \quad \textit{(partial moderation)},\\
\theta = \widetilde{\theta}(\rho) &\implies y^*(\theta,\rho) = \rho \quad \textit{(balanced)},\\
\theta > \widetilde{\theta}(\rho) &\implies 0 < y^*(\theta,\rho) < \rho \quad \textit{(partial polarization)},\\
\theta \to 1 &\implies y^*(\theta,\rho) \to \rho \quad \textit{(balanced)}.
\end{aligned}$$

(iii) If $\rho \geq \sqrt{5}-2$, then there exists a threshold $\widehat{\theta}(\rho)$ such that the stable steady state corresponds to full moderation, that is, $y^*(\theta, \rho) = 1$ for all $\theta \leq \widehat{\theta}(\rho)$; decreases thereafter for $\theta > \widehat{\theta}(\rho)$ (partial moderation); and converges to the balanced steady-state ρ as the network gets infinitely dense, that is, $\lim_{\theta \to 1} y^*(\theta) = \rho$.

Proposition 3 provides a rich bifurcation structure of the stable steady-state behavior as a function of the network density parameter θ. We see that in all cases sparse networks result in full moderation and dense networks result in balanced steady-states. When the preference advantage is not sufficiently in favor of the extremes (part (iii)), then, increasing network density weakens the absolute strength of each source of advantage and there is a monotonic transition to a balanced steady-state as the network becomes dense. When preference advantage is sufficiently strong (parts (i) and (ii)), then, the transition is non-monotonic and includes a range of densities where partial polarization arises. Finally, when the preference advantage is sufficiently large (part (i)), there will also be a range of densities where full polarization arises.

Figure 2 illustrates these three cases and the threshold case $\rho = \frac{1}{9}$. Together, the panels in Figure 2 underscore the non-monotonic relationship between connectivity and content that arises when the preferences in the population are sufficiently extreme.

4.4 Impact of a mean-preserving spread

We now examine how a mean-preserving spread of the degree distribution affects the thresholds for full moderation ρ_0 and full polarization ρ_1 in Proposition 1.

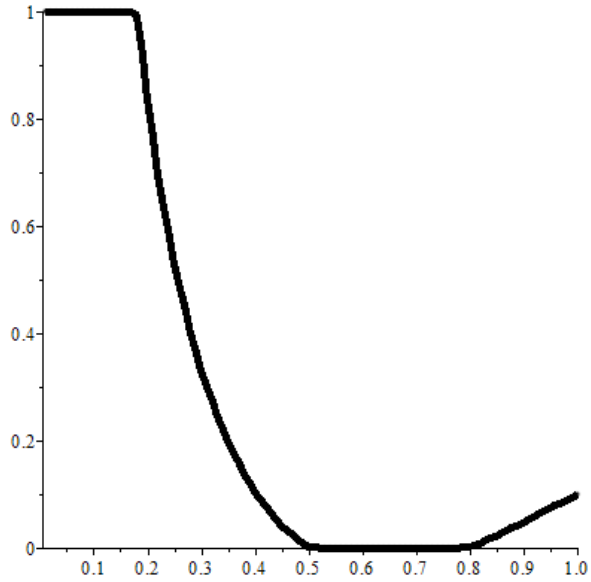
Proposition 4. *There exists some $\epsilon > 0$, a mean-preserving spread $\{p'_k\}$ of the degree distribution $\{p_k\}$ in which $p_1 = p'_1$, and a mean preserving spread $\{p''_k\}$ of $\{p'_k\}$ where $p''_1 < p'_1 + \epsilon$, such that:*

(i) ρ_0 decreases under both $\{p'_k\}$ and $\{p''_k\}$; and

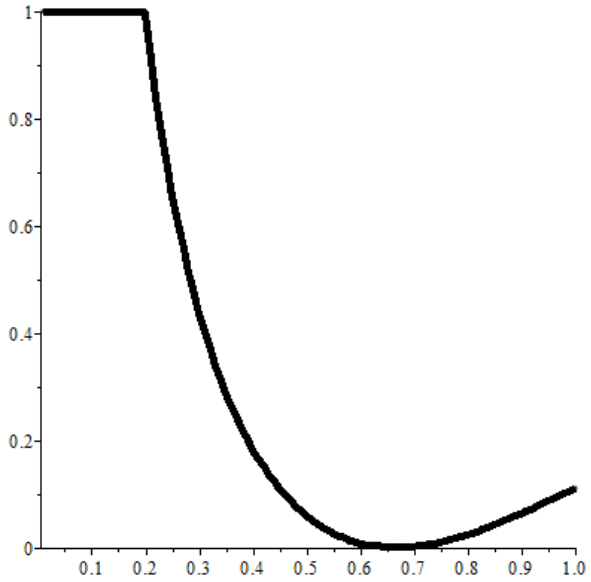
(ii) ρ_1 is unchanged under $\{p'_k\}$ and increases under $\{p''_k\}$.

Note that both p'_k and p''_k are mean-preserving spreads of the original distribution p_k. The results in Proposition 4 show that increasing the dispersion of the degree distribution—via mean-preserving spreads (subject to not increasing the probability of just one connection by too much)—will expand the intermediate region of the parameter space that supports *partial moderation and polarization.*

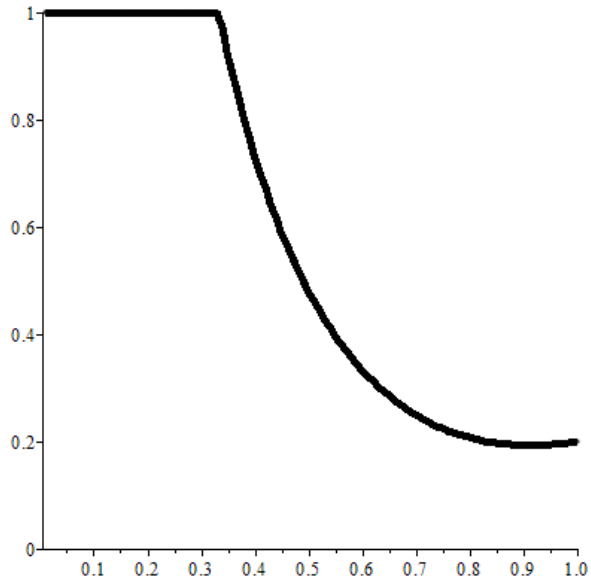
Mean-preserving spreads generate a more heterogeneous social environment, characterized by a larger number of highly connected individuals (potential influencers) and marginally connected individuals. This increased heterogeneity dampens the tendency of content filtering to converge toward the extreme outcomes of full moderation or full polarization.



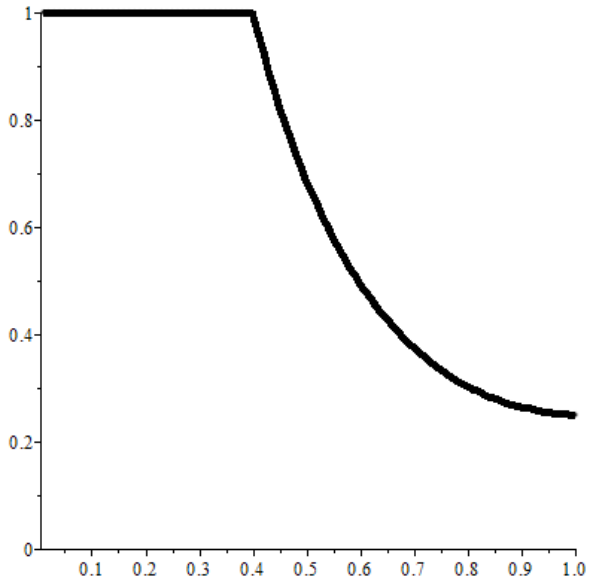
(a) $\rho = 0.1$



(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

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 $1 - s^* > 0$.

Proposition D1 in Online Appendix D.1 supports the robustness of our earlier findings: namely, that polarization does not arise under high ρ, while full polarization emerges under low ρ. It shows that, with the option of remaining silent, polarization becomes less likely to emerge than in the

baseline model.

Proposition D2 in Online Appendix D.1, which can be viewed as a counterpart to Proposition 2, characterizes how network density affects the degree of polarization. This proposition shows that allowing for silence preserves the key non-monotonic relationship between network density and polarization established in the baseline model.

Proposition D3 in Online Appendix D.1 refines Proposition D2 by fully characterizing the unique stable equilibrium under an exponential degree distribution. Figure D4 in the Online Appendix D.1 illustrates this characterization for $\theta = 0.8$. The proposition reveals a key structural insight: silence functions as a behavioral *filter*. Instead of compelling agents to recommend misaligned content, it provides an exit option, thereby dampening polarization. More generally, Proposition D3 and Figure D4 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	$\frac{1-\rho}{2} + \epsilon$

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:

$$\ell_t = \left(\frac{1-\rho}{2} - \epsilon\right)(1 - G(1 - \ell_{t-1})) + \rho \frac{\ell_{t-1}}{\ell_{t-1} + r_{t-1}} G(\ell_{t-1} + r_{t-1}) + \left(\frac{1-\rho}{2} + \epsilon\right) G(\ell_{t-1}), \quad (17)$$

$$r_t = \left(\frac{1-\rho}{2} + \epsilon\right)(1 - G(1 - r_{t-1})) + \rho \frac{r_{t-1}}{\ell_{t-1} + r_{t-1}} G(\ell_{t-1} + r_{t-1}) + \left(\frac{1-\rho}{2} - \epsilon\right) G(r_{t-1}). \quad (18)$$

The second terms of the right-hand side of equations (17) and (18), 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.

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 D4 in Online Appendix D.2 characterizes how the steady-state equilibrium of the system depends on ρ, the share of type M-agents, 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 the non-polarized equilibrium and those sustaining the polarized equilibrium become narrower.

Overall, Proposition D4 demonstrates that the range of ρ that supports full polarization or full moderation 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 non-polarized equilibrium or total polarization.

6 Concluding remarks

We have developed a simple model of content filtering on social media that illustrates how the structure of social networks influences the diffusion of different types of ideological content. The model yields rich behavior and offers insights into how connectivity fosters moderate versus extreme content. We highlight the influence of a preference advantage and pairwise comparison advantage in our model.

Our results show that the density of connections among individuals significantly influences the steady state of the system. Remarkably, regardless of the population's preference distribution, the pairwise comparison advantage dominates in sparse networks resulting in full moderation. At the

opposite extreme, dense networks dampen both sources of advantage, and the steady state aligns with the population's proportion of each type. Between these extremes, more complex effects arise when preference and pairwise comparison advantage counterbalance each other. In particular, we observe that the transition can (i) be non-monotonic, passing through partially and potentially fully polarized steady states, and (ii) exhibit rapid shifts between a fully moderate and a fully polarized steady state when type M-agents are scarce and the network remains relatively sparse.

Our findings compliment the existing empirical research on social media's influence on recent trends of populism and polarization. In contrast to this literature we show that even when the composition of these connections is held constant in a population, but the frequency increases then the process of content filtering may fundamentally shape the information environment, influencing the emergence of moderate versus extreme content.

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. 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 research is to endogenize the network formation process. Agents could strategically form connections based on ideological proximity or content alignment, giving rise to endogenous echo chambers or ideological segregation. While such an extension would offer deeper insights into the co-evolution of social fragmentation and polarization, it would also entail greater analytical complexity.

Another potential extension is to incorporate homophily into the model. Specifically, with probability α, agents would receive a recommendation from someone of the same type,[14] and with probability $1 - \alpha$, from a randomly selected individual in the population. This framework would allow us to examine how varying the degree of homophily (α) influences equilibrium outcomes.[15]

We leave these exciting avenues for future research.

[14]This could be generated by a social media algorithm, which analyzes user behavior and interactions to determine which content is most relevant and engaging for each individual user.

[15]For instance, Enikolopov et al. (2024) find that homophilic connections leads to increased social media usage, which increases polarization, as individuals became less connected across income strata and less likely to share the same political opinions with others.

Appendix: 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:

$$f_y(y,\rho) = \rho G'(1-y) + (1-\rho)\left[\frac{1}{2}G'\left(\frac{1+y}{2}\right) + \frac{1}{2}G'\left(\frac{1-y}{2}\right)\right]. \tag{19}$$

Given that G' is positive, $f_y(y,\rho)$ is also positive, which proves (i).

(ii) Equation (19) 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:

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

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

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

Hence, $f(y,\rho) - y > 0$ in the vicinity of $y = 0$, and $f(y,\rho) - y < 0$ 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 $f(y,\rho)$ only has one inflection point, the interior solution y^* is unique. Finally, we have:

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

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

Hence, it must be that

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

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

Next, we 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:

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

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

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

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, $f_\rho(y,\rho) > 0 \Longrightarrow \frac{dy^*(\rho)}{d\rho} > 0$.

It remains to prove that equation $y^*(\rho) = \rho$ has a unique interior solution $\widetilde{\rho} \in (\rho_0,\rho_1)$, such that $\widetilde{\rho} < \frac{1}{3}$ and $y^*(\rho) \lesseqgtr \rho \Longleftrightarrow \rho \lesseqgtr \widetilde{\rho}$. From $\frac{dy^*(\rho)}{d\rho} > 0$ for every $\rho \in (\rho_0,\rho_1)$, and from

$$\lim_{\rho\to\rho_0} y^*(\rho) = 0, \qquad \lim_{\rho\to\rho_1} y^*(\rho) = 1.$$

we infer that $\rho \to y^*(\rho)$ is a bijective mapping which maps (ρ_0,ρ_1) onto $(0,1)$. Hence, it has a single-valued inverse mapping $y \to \varrho(y)$, which maps $(0,1)$ onto (ρ_0,ρ_1). This inverse mapping

can be written explicitly by solving (5) w.r.t. ρ:

$$\rho = \varrho(y) := \frac{y - \left[G\left(\frac{1+y}{2}\right) - G\left(\frac{1-y}{2}\right)\right]}{1 - G(1-y) - \left[G\left(\frac{1+y}{2}\right) - G\left(\frac{1-y}{2}\right)\right]}. \tag{20}$$

Using the l'Hospital's rule, it is readily verified that $\varrho(y)$ satisfies

$$\lim_{y\to 0} \varrho(y) = \rho_0, \qquad \lim_{y\to 1} \varrho(y) = \rho_1.$$

Also, from $\varrho(\cdot) = y^{*-1}(\cdot)$ and $y^{*\prime}(\cdot) > 0$, it follows that $\varrho'(\cdot) > 0$. Thus, the unique stable steady state $y^*(\rho)$ can be written as follows:

$$y^*(\rho) = \begin{cases} 0, & \rho \le \rho_0; \\ \varrho^{-1}(\rho), & \rho_0 < \rho < \rho_1; \\ 1, & \rho > \rho_1. \end{cases}$$

In geometric terms, existence and uniqueness of $\widetilde{\rho}$ means that the curve $y = \varrho^{-1}(\rho)$ intersects only once the 45°-line on the (ρ, y)-plane. Or, equivalently, that the curve given by $\rho = \varrho(y)$ intersects only once the 45-degree line on the (y, ρ)-plane, i.e., that the equation $y = \varrho(y)$ has a unique solution. Using the definition (20) of $\varrho(y)$, one can simplify the equation $y = \varrho(y)$ to:

$$\frac{G(1-y)}{1-y} = \frac{G\left(\frac{1+y}{2}\right) - G\left(\frac{1-y}{2}\right)}{y}. \tag{21}$$

Because $G(\cdot)$ is a convex function, the LHS of (21), $\frac{G(1-y)}{1-y}$, increases w.r.t. $1-y$, hence decreases w.r.t. y. Similarly, because numerator of the RHS of (21) is convex in y, the RHS increases w.r.t. y. Therefore, (21) has a unique solution $\Longrightarrow$ $y = \varrho(y)$ has a unique solution $\Longrightarrow$ $y^*(\rho) = \rho$ has a unique interior solution $\widetilde{\rho} \in (\rho_0, \rho_1)$. That $y^*(\rho) \lesseqgtr \rho \iff \rho \lesseqgtr \widetilde{\rho}$ follows from combining the uniqueness of $\widetilde{\rho}$ with the following inequalities:

$$0 = y^*(\rho_0) < \rho_0 < \widetilde{\rho} < \rho_1 < y^*(\rho_1) = 1.$$

It remains to prove that $\widetilde{\rho} < \frac{1}{3}$, or, equivalently, that $y^*\left(\widetilde{\rho}\right) < \frac{1}{3}$. Because $G(\cdot)$ is convex, the inequality

$$\frac{G\left(c\right) - G(a)}{c-a} < \frac{G\left(c\right) - G\left(b\right)}{c-b},$$

holds for any a, b, c satisfying $0 \le a < b < c \le 1$. By setting $a = 0$, $b = \frac{1}{3}$, and $c = \frac{2}{3}$, we get:

$$\frac{G\left(\frac{2}{3}\right) - G(0)}{\frac{2}{3} - 0} < \frac{G\left(\frac{2}{3}\right) - G\left(\frac{1}{3}\right)}{\frac{2}{3} - \frac{1}{3}},$$

which can be equivalently rewritten as

$$\frac{G\left(1-\frac{1}{3}\right)}{1-\frac{1}{3}} < \frac{G\left(\frac{1+\frac{1}{3}}{2}\right) - G\left(\frac{1-\frac{1}{3}}{2}\right)}{\frac{1}{3}}. \tag{22}$$

Observe that the LHS/RHS of (22)) is the LHS/RHS of (21) evaluated at $y = \frac{1}{3}$. Since $y^*(\widetilde{\rho})$ is the unique solution to (21), and the LHS/RHS of (21) decreases/increases, it follows that $y^*(\widetilde{\rho}) < \frac{1}{3}$, hence $\widetilde{\rho} < \frac{1}{3}$. This completes the proof. □

Proof of Proposition 2

We will need the following Lemma.

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

$$\lim_{\theta \to 0} \rho_1(\theta) = 0 \tag{23}$$

$$\lim_{\theta \to \overline{\theta}} \rho_1(\theta) = 1. \tag{24}$$

Proof of Lemma 2. From Assumption 2, $\lim_{\theta \to \overline{\theta}} \mu(\theta) \to \infty$, hence (24) follows immediately from (7). Next, from Assumption 3, we get, under $\theta \to 0$

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

which implies (23).

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

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

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

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

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

$$\frac{\mu'(\theta)}{\mu(\theta)-1} + \frac{p_1'(\theta)}{1-p_1(\theta)} = \sum_{k=1}^{\infty} \frac{p_k'(\theta)}{p_k(\theta)} a_k(\theta), \tag{25}$$

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

$$a_k(\theta) := \begin{cases} 0, & k = 1; \\ \left(\frac{k-1}{\mu(\theta)-1} - \frac{1}{1-p_1(\theta)}\right) p_k(\theta), & k \geq 2. \end{cases} \tag{26}$$

Representation (25) – (26) is readily verified using the following identities:

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

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

where δ_{1k} is the Kronecker's delta. The sequence $\{a_k(\theta)\}$ defined by (26) 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 (25) as follows:

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

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. □

We now proceed with the proof of Proposition 2.

(*i*) From Lemma 2, $\rho_1(\theta)$ increases from 0 to 1 as θ increases from 0 to $\overline{\theta}$. Hence, the equation $\rho_1(\theta) = \rho$ has the unique solution $\widehat{\theta}(\rho)$ w.r.t. θ for each $\rho \in (0,1)$, and

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

Also,

$$\lim_{\theta\to 0}\rho_1(\theta)=0 \quad\Longrightarrow\quad \lim_{\rho\to 0}\widehat{\theta}(\rho)=0.$$

This proves (i).

(ii) 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 $\lim_{\theta\to\overline{\theta}}\mu(\theta)=\infty$. Define[16]

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

If $\rho<\widehat{\rho}$, the equation $\rho_0(\theta)=\rho$ has at least two distinct roots. Let $\theta_1(\rho)$ and $\theta_2(\rho)>\theta_1(\rho)$ be, respectively, the least and the second least of these roots. Clearly, $\theta_1(\rho)>\widehat{\theta}(\rho)$, as $\rho_1(\theta)>\rho_0(\theta)$. Furthermore, $\rho<\rho_0(\theta)\,\forall\,\theta\in(\theta_1(\rho),\theta_2(\rho))$. Combinig this with Proposition $1(i)$, we get:

$$\theta\in(\theta_1(\rho),\theta_2(\rho)) \quad\Longrightarrow\quad y^*(\theta)=0,$$

which means full polarization. Also,

$$\lim_{\theta\to 0}\rho_0(\theta)=0 \quad\Longrightarrow\quad \lim_{\rho\to 0}\theta_1(\rho)=0;$$

$$\lim_{\theta\to\overline{\theta}}\rho_0(\theta)=0 \quad\Longrightarrow\quad \lim_{\rho\to 0}\theta_2(\rho)=\overline{\theta}.$$

It remains to prove that $\widehat{\rho}<1/3$. By Proposition 1, there is a threshold $\widetilde{\rho}(\theta)$ separating polarization and moderation, such that $1/3>\widetilde{\rho}(\theta)>\rho_0(\theta)\ \forall\,\theta\in(0,\overline{\theta})$. Hence, $\widehat{\rho}<1/3$. This proves (ii).[17]

(iii) The result follows immediately if we set $\theta\to\overline{\theta}$ and notice that, from Assumption 2, $\lim_{\theta\to\overline{\theta}}G(x,\theta)=0\ \forall\,x\in(0,1)$. Hence the RHS in (5) goes to $\rho\ \forall y\in(0,1)$. This proves (iii) and completes the proof. □

Proof of Proposition 3

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

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

where $\theta\in(0,1)$ is the density parameter. 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

[16]Strictly speaking, this definition of $\widehat{\rho}$ only makes sense if $\overline{\theta}<\infty$. If $\overline{\theta}=\infty$, one should define $\widehat{\rho}$ as the maximum of $\rho_0(\theta)$ over a sufficiently large compact interval $\Theta\subset\mathbb{R}_+$, such that $\rho_0(\theta)$ is sufficiently small beyond that interval. Such an interval always exists since $\lim_{\theta\to 0}\rho_0(\theta)=\lim_{\theta\to\overline{\theta}}\rho_0(\theta)=0$.

[17]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.

equilibrium $y^*(\theta,\rho)$ is given by

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

where

$$\widetilde{y}(\theta,\rho) := \frac{\sqrt{(1-\rho-2\theta)^2 + 16\rho(1-\theta)} - (1-\rho)}{2\theta}. \tag{27}$$

The function $\widetilde{y}(\theta,\rho)$ defined by (27) is quasiconvex w.r.t. θ over $(0,1)$. To see this, observe that the equation $\widetilde{y}(\theta,\rho) = y$ cannot have more than two solutions in $(0,1)$ for any $y \in \mathbb{R}$. Indeed, using (27), the equation $\widetilde{y}(\theta,\rho) = y$ can be equivalently restated as

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

which is a quadratic equation w.r.t. θ, hence it has at most two real solutions. Furthermore, by setting $\theta \to 0$ and $\theta \to 1$ in the RHS of (27), 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 w.r.t. 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) Assume $\rho < \frac{1}{9}$. Then, from (27), equation $\widetilde{y}(\theta,\rho) = 0$ has two distinct roots, $0 < \theta_1(\rho) < \theta_2(\rho) < 1$, which can be expressed in closed form as

$$\theta_{1,2}(\rho) = \frac{1}{2} + \frac{3}{2}\left[\rho \mp \sqrt{(1-\rho)\left(\frac{1}{9} - \rho\right)}\right].$$

It then follows from the quasi-convexity of $\widetilde{y}(\theta,\rho)$ that

$$y^*(\theta,\rho) = 0 \quad \Longleftrightarrow \quad \theta_1(\rho) \le \theta \le \theta_2(\rho).$$

Furthermore, one can show that equation $\widetilde{y}(\theta,\rho) = \rho$ has two solutions w.r.t. θ. One of those is $\overline{\theta} = 1$, which implies $\lim_{\theta\to 1} y^*(\theta,\rho) = \rho$, while the other solution is interior and is given by

$$\widetilde{\theta}(\rho) := \underbrace{\frac{4\rho}{1-\rho^2} \in (0,1)}_{\text{because } \rho < \frac{1}{9}}. \tag{28}$$

Thus, from the quasi-convexity of $\widetilde{y}(\theta,\rho)$, we have:

$$y^*(\theta,\rho) \gtreqless \rho \quad \Longleftrightarrow \quad \theta \lesseqgtr \widetilde{\theta}(\rho).$$

Finally, let us define $\widehat{\theta}(\rho)$ as the unique solution of the equation $\widetilde{y}(\theta,\rho)=1$ w.r.t. θ:

$$\widehat{\theta}(\rho) := \frac{2\rho}{1+\rho}. \tag{29}$$

From the quasi-concavity of $\widetilde{y}(\theta,\rho)$, it follows that

$$y^*(\theta,\rho)=1 \quad \Longleftrightarrow \quad \theta \leq \widehat{\theta}(\rho).$$

Collecting the results obtained above and using quasi-convexity of $\widetilde{y}(\theta,\rho)$ proves (i).

(ii) Assume now that $\frac{1}{9}<\rho<\sqrt{5}-2$. Then, $\widetilde{y}(\theta,\rho)>0\ \forall\ \theta(0,1)$. 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, hence it varies non-monotonically.[18] The thresholds $\widetilde{\theta}(\rho)$ and $\widehat{\theta}(\rho)$ are given by (28) – (29). Note that, from (28),

$$\widetilde{\theta}(\rho)<1 \quad \Longleftrightarrow \quad \rho<\sqrt{5}-2.$$

This proves (ii).

(iii) Finally, assume that $\rho\geq\sqrt{5}-2$. In this case, $\widetilde{y}(\theta,\rho)>0\ \forall\theta(0,1)$. 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, as θ varies from $\widehat{\theta}(\rho)$ to $\overline{\theta}=1$. The threshold $\widehat{\theta}(\rho)$ is given by (29). This proves (iii) and completes the proof. □

Proof of Proposition 4

First, from (6) one can immediately observe that ρ_0 is decreasing in $\psi=\mathbb{E}[k(1/2)^{k-1}]$. The expression $k(1/2)^{k-1}$ is convex for $k\geq 2$, hence, the mean-preserving spread $\{p'_k\}$ (where $p'_1=p_1$) leads to an increase in ψ, hence to a decrease in ρ_0. By continuity, one can also construct a mean-preserving spread $\{p''_k\}$ of $\{p'_k\}$ that has the same effects provided p''_1 does not increase by too much ($<\epsilon$). Second, one can readily observe from (7) that ρ_1 is increasing in p_1 and is otherwise unaffected by a mean-preserving spread. Hence, any mean-preserving spread that increases the value of p_1 will increase ρ_1. This completes the proof. □

[18] When $\rho=\frac{1}{9}$, $\widetilde{y}(\theta,\rho)=0$ at $\theta=\frac{2}{3}$, which is the minimizer of $\widetilde{y}(\theta,\rho)$ and its tangency point with the horizontal axis. Thus, as $\rho\nearrow\frac{1}{9}$, the interval of full polarization shrinks to a single point.

References

Algan, Y., N. Dalvit, Q.-A. Do, A. Le Chapelain, and Y. Zenou (2025). Friendship networks and political opinions. *American Economic Review*.

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.

Arora, S. D., G. P. Singh, A. Chakraborty, and M. Maity (2022). Polarization and social media: A systematic review and research agenda. *Technological Forecasting and Social Change 183*, 121942.

Bakshy, E., S. Messing, and L. A. Adamic (2015). Exposure to ideologically diverse news and opinion on facebook. *Science 348*(6239), 1130–1132.

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 176*, 105045.

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. (2015). Word of mouth model of sales. *Economics Letters 133*, 45–50.

Campbell, A. (2019). Social learning with differentiated products. *The RAND Journal of Economics 50*(1), 226–248.

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*.

Enikolopov, R., M. Petrova, G. Russo, and D. Yanagizawa-Drott (2024). Socializing alone: How online homophily has undermined social cohesion in the US. Unpublished manuscript, University of Zurich.

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

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.

Messing, S. and S. J. Westwood (2014). Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication Research 41*(8), 1042–1063.

Mueller, B. (2025). Does social media cause polarization? A cross-country analysis using staggered smartphone adoption. Available at SSRN: https://ssrn.com/abstract=5327921.

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.

Nyhan, B., J. Settle, E. Thorson, M. Wojcieszak, P. Barberá, A. Y. Chen, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Dimmery, et al. (2023). Like-minded sources on facebook are prevalent but not polarizing. *Nature 620*(7972), 137–144.

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

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.

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

Online Appendix: Additional results and formal propositions for the extensions

A Verifying $\rho_0 \leq \rho_1$

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

$$(\mu + p_1 - 2)(\mu - 1) \geq 2(1 - \psi)(1 - p_1).$$

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

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

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 $\mu < \infty$, the following inequality holds*

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

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:

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

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

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

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

holds for all $k \in \mathbb{N}$, with ">" for $k \geq 3$. 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

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

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 $\rho_1=\frac{\mu+p_1-2}{\mu+p_1-2p_1}$, since $\mu+p_1$ 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. In particular, we can construct these FOSD-respecting shocks so that:

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

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 Extensions: Formal results

D.1 Allowing for no recommendation

In this section, we derive a series of formal results when we allow for no recommendation.

D.1.1 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 (D.3) takes the form

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

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. □

D.1.2 Other results

Proposition D1. *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, $\rho_0^{\text{silent}} < \rho_0$.

Proof of Proposition D1

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

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

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

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

$$\mathbf{J}(y^*, s^*) = \begin{pmatrix} \rho G'(1-y^*) + \frac{1-\rho}{2}\left(G'\left(\frac{1+y^*+s^*}{2}\right) + G'\left(\frac{1-y^*+s^*}{2}\right)\right) & \frac{1-\rho}{2}\left(G'\left(\frac{1+y^*+s^*}{2}\right) - G'\left(\frac{1-y^*+s^*}{2}\right)\right) \\ -\frac{1-\rho}{2}G'\left(\frac{1-y^*+s^*}{2}\right) & \rho\, G'(s^*) + \frac{1-\rho}{2}\, G'\left(\frac{1-y^*+s^*}{2}\right) \end{pmatrix}. \tag{D.3}$$

(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 (D.3) for the Jacobian evaluated at the steady state, we get the Jacobian $\mathbf{J}(1, 0)$ evaluated at the non-polarized equilibrium:

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

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

$$\operatorname{tr}(\mathbf{J}(1,0)) = \rho p_1 + \frac{1-\rho}{2}\mu + p_1;$$

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

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

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

$$\lambda_2 = p_1.$$

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

$$\rho > \rho_1^{\text{silent}} := \begin{cases} 0, & \mu \leq 2; \\ \frac{\frac{\mu}{2}-1}{\frac{\mu}{2}-p_1}, & \mu > 2. \end{cases} \tag{D.4}$$

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 $y^* = 0$:

$$s = \rho G(s) + (1-\rho) G\left(\frac{1+s}{2}\right). \tag{D.5}$$

Because the RHS of (D.5) is an increasing and convex function of s, positive at $s = 0$ and equal to one at $s = 1$, (D.5) 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,

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

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 (D.3) 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

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

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. □

Proposition D2. *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:

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

Proof of Proposition D2

(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 D1(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:

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

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 $\mu(\theta) = G_x(1,\theta) > 2$. 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: proving that $\widehat{\rho} > 0$. As the denominator in $\frac{1-G_x\left(\frac{1+\overline{s}(\theta)}{2},\theta\right)}{\mu(\theta)-G_x\left(\frac{1+\overline{s}(\theta)}{2},\theta\right)}$ 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:

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

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

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

Such θ_0 exists because $\frac{1-G_x\left(\frac{1+\overline{s}(\theta)}{2}, \theta\right)}{\mu(\theta)-G_x\left(\frac{1+\overline{s}(\theta)}{2}, \theta\right)}$ 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

$$\rho G'_x(1, \theta) + (1 - \rho) G'_x\left(\frac{1 + s^*\,(\rho, \theta)}{2}, \theta\right) < 1 \tag{D.6}$$

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. □

Proposition D2 shows that 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. 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.

Interior equilibria are generally difficult to characterize analytically under an arbitrary degree

distribution. To gain sharper insights, in Proposition D3, we 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.

Proposition D3. *Define* $\rho_0^{\text{silent}}(\theta)$ *and* $\rho_1^{\text{silent}}(\theta)$ *as follows:*

$$\rho_0^{\text{silent}}(\theta) := \begin{cases} 0, & \theta \le 0.75; \\ \textit{solves } \rho\dfrac{1}{1-\theta} + (1-\rho)\dfrac{16(1-\theta)}{(2-\theta+\sqrt{4\rho-4\rho\theta+\theta^2})^2} = 1, & \theta > 0.75; \end{cases} \tag{D.7}$$

$$\rho_1^{\text{silent}}(\theta) := \begin{cases} 0, & \theta \le 0.5; \\ \dfrac{2\theta-1}{4\theta-(2\theta^2+1)}, & \theta > 0.5. \end{cases} \tag{D.8}$$

Then, 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 $\rho \le \rho_0^{\text{silent}}(\theta) < \rho_0(\theta)$*, the unique stable equilibrium exhibits full polarization with a positive fraction of agents being silent, 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 $\rho \ge \rho_1^{\text{silent}}(\theta)$*, the unique stable equilibrium exhibits full moderation 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:*

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

Proof of Proposition D3.

(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

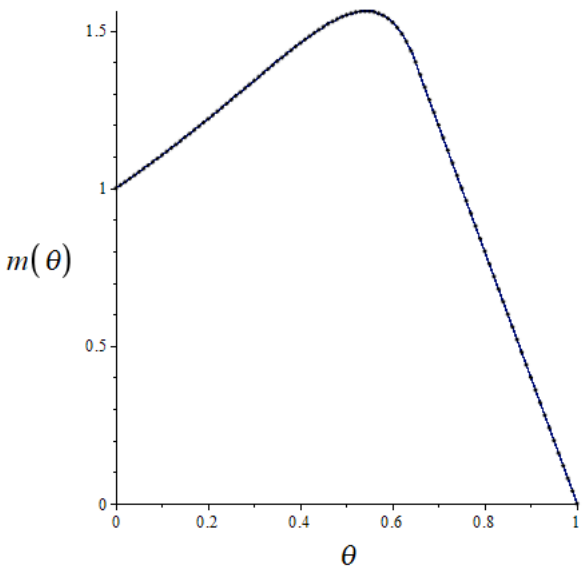


Figure D1: How $m(\theta) := \min_{\rho\in[0,1]} \phi(\rho,\theta)$ varies with θ.

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 (D.6) for stability of polarized equilibrium becomes

$$\phi(\rho,\theta) < 1, \tag{D.9}$$

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

$$\phi(\rho,\theta) := \rho\frac{1}{1-\theta} + (1-\rho)\frac{16\,(1-\theta)}{\left(2-\theta+\sqrt{4\rho-4\rho\theta+\theta^2}\right)^2}. \tag{D.10}$$

We need the following Lemma.

Lemma D2. *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 D2. *(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 D1. This proves *(a)*.

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

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

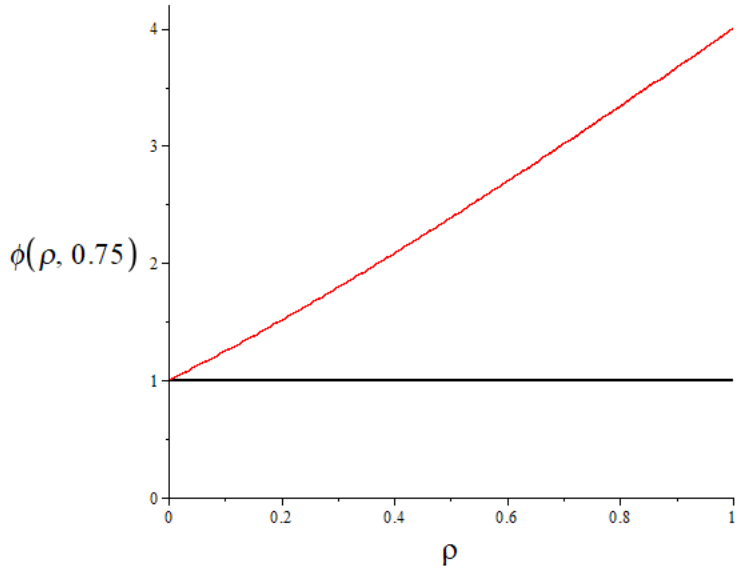


Figure D2: How $\phi(\rho, 0.75)$ varies with ρ.

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

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 D2).

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 D2. □

We now proceed with the proof of Proposition D3(i). From Lemma D2(a), the stability condition (D.9) or the polarized equilibrium fails to hold when for $\theta \leq 0.75$. From Lemma D2(b), (D.9) 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:

$$s_t \gtreqless s_{t-1} \Longleftrightarrow y_{t-1} \lesseqgtr \tilde{y}\left(s_{t-1}\right), \tag{D.11}$$

where

$$\tilde{y}(s) := 1 + s - \frac{2s\left(1-\rho+\rho\theta\right) - 2\theta s^2}{\left(1-\rho\right)\left(1-\theta\right) + \theta^2 s - \theta^2 s^2}. \tag{D.12}$$

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

have:

$$y_t \gtreqless y_{t-1} \Longleftrightarrow s_{t-1} \gtreqless \widetilde{s}(y_{t-1}), \tag{D.13}$$

where

$$\widetilde{s}(y) := \frac{2}{\theta}\left[1 - \sqrt{\left(\frac{1}{2}\theta y\right)^2 + (1-\rho)(1-\theta)\frac{1-\theta+\theta y}{(1-\theta+\theta y)-\rho}}\right] - 1. \tag{D.14}$$

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

$$y = \widetilde{y}(s), \qquad s = \widetilde{s}(y), \tag{D.15}$$

where the functions $\widetilde{y}(s)$ and $\widetilde{s}(y)$ are defined by, respectively, (D.12) and (D.14). From (D.11) and (D.13), it is clear that the two curves in (D.15) correspond to the loci of, respectively, $y_t = y_{t-1}$ and $s_t = s_{t-1}$. Furthermore, one can show that the function $\widetilde{y}(s)$ 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:

$$\frac{1-\frac{1}{2}\theta}{\frac{1}{2}\theta - (1-\theta)\frac{\rho}{1-\rho}} < \frac{1}{1-\theta}\left(1 + \theta + 2\theta\frac{\rho}{1-\rho}\right). \tag{D.16}$$

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

$$\widetilde{s}(0) < s^*(\theta, \rho) = \frac{2 - \theta - \sqrt{4\rho - 4\rho\theta + \theta^2}}{2\theta}, \tag{D.17}$$

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 (D.16) and (D.17) 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 D4c. The directions of the dynamics are determined by (D.13) and (D.11). This proves (ii).

(iii) The expression $\rho_1^{\text{silent}}(\theta) = \frac{2\theta-1}{4\theta-2\theta^2-1}$ for the exponential degree distribution can be obtained by plugging $p_1(\theta) = 1 - \theta$ and $\mu(\theta) = \frac{1}{1-\theta}$ into (D.8). Thus, (iii) follows immediately from Proposition D1(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 D3.

From Figure D3, one can see that the stability condition (D.9) holds for a non-empty set of

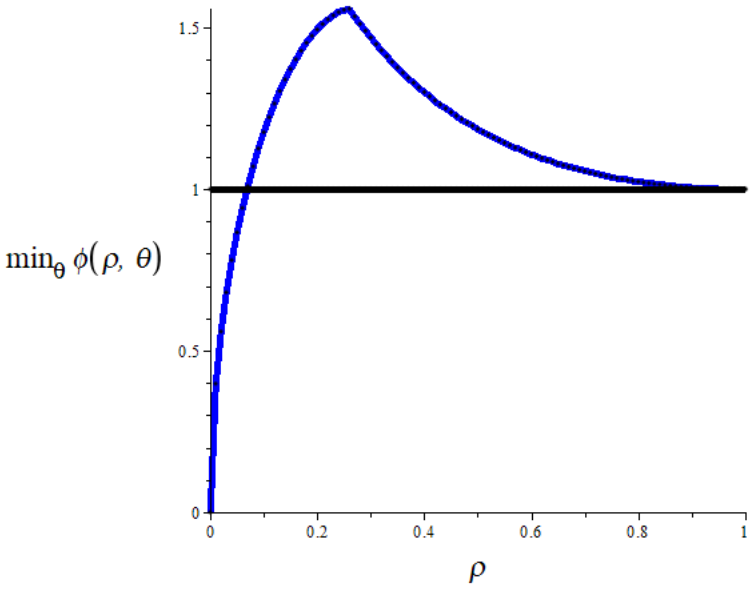


Figure D3: How $\min_\theta \phi(\rho, \theta)$ varies with ρ.

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 D1 and D2. This proves (iv) and completes the proof. □

Proposition D3 shows that, 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. 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 networks, the abundance of centrist contents ensures that agents repeatedly encounter moderate content and continue to spread it.

When θ is large, the threshold $\rho_0^{\text{silent}}(\theta)$ is a decreasing function 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 suppresses the reinforcement of extreme views when centrists

are scarce.

Figure D4 below illustrates Proposition D3 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.[1] Panel D4a shows that, when $\rho > \rho_1^{\text{silent}}(\theta) = \frac{15}{23}$, the unique stable equilibrium is the non-polarized equilibrium ($y^* = 1$). Panel D4b shows that, when $\rho < \rho_0^{\text{silent}}(\theta) \approx 0.0522$, the unique stable equilibrium is the polarized equilibrium ($y^* = 0$). Finally, panel D4c 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 D4 also provides phase diagrams which allow one establish qualitatively the stability of the non-polarized equilibrium on panel D4a, the polarized equilibrium on panel D4b, and the interior equilibrium on panel D4c. 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 (D.11) and (D.13) in the Appendix.

D.2 Allowing for asymmetry

First, observe that the expression

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

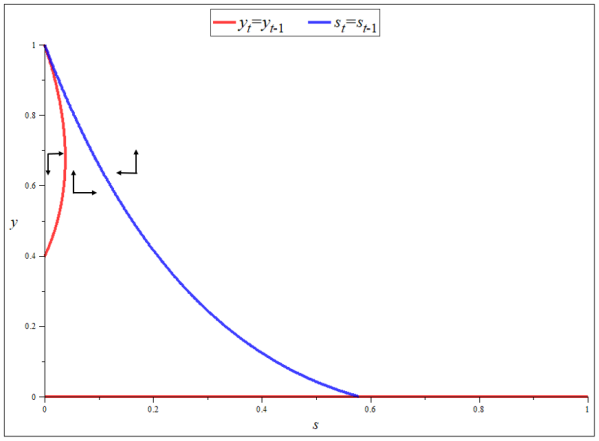
can be derived algebraically as follows:

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

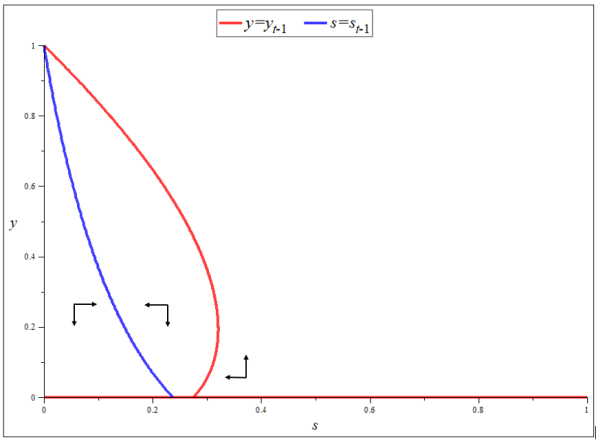
where the last line gives the full expression for the probability that an M-type agent randomly adopts an L-type recommendation when exposed only to extreme content.

Proposition D4. *We obtain the following results for a right-skewed preference distribution ($\epsilon > 0$):*

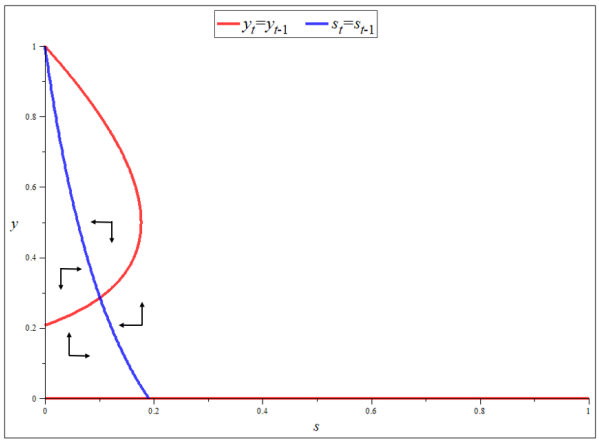
[1] Technically, there is one more steady state—the "full-silence" steady state, in which $s^* = 1$. However, we show in Online Appendix D.1.1 that this steady state is never stable, therefore we disregard it as uninteresting.



(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 D4: Equilibrium characterization under exponential distribution with $\theta = 0.8$.

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

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

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 $(\ell^*(\epsilon), r^*(\epsilon))$.

Proof of Proposition D4.

(i) The steady-state conditions of the two-dimensional dynamical system (17)-(18) are given by:

$$\ell = \left(\frac{1-\rho}{2} - \epsilon\right)(1 - G(1-\ell)) + \rho\frac{\ell}{\ell + r}G(\ell + r) + \left(\frac{1-\rho}{2} + \epsilon\right)G(\ell); \tag{D.18}$$

$$r = \left(\frac{1-\rho}{2} + \epsilon\right)(1 - G(1-r)) + \rho\frac{r}{\ell + r}G(\ell + r) + \left(\frac{1-\rho}{2} - \epsilon\right)G(r). \tag{D.19}$$

It is readily verified that the non-polarized equilibrium $(\ell^*, r^*) = (0, 0)$ satisfies (D.18)-(D.19). The Jacobian of the dynamical system (17)-(18) 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:

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

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:

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

This proves (i).

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

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

Hence, the steady-state conditions (D.18) – (D.19) 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

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

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

The Jacobian of the dynamical system (17)-(18) evaluated at the polarized equilibrium is given by:

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

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:

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

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:

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

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

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

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

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

Hence,

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

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

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

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 $\lambda_1(\rho,\epsilon) = 1$, decreases in response to a bit more asymmetry. This proves (ii) and completes the proof. □

Part (i) of Proposition D4 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. 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.