Article

Communication Research
1–32
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00936502251362278
journals.sagepub.com/home/crx
Sage

When Opinions Polarize Without Persuasion: Modeling the Dynamics of Attitude-Opinion Convergence and Decoupling

Dongyoung Sohn[1]

Abstract

Digital media is often blamed for deepening societal divides by fostering echo-chambers that reinforce biases. However, the polarized opinions visible on the media may not necessarily indicate deeper fragmentation of hidden beliefs, which is often assumed to be driven by persuasion. Instead, public opinion polarization can emerge from contextual dynamics that decouple private attitudes from expressed opinions. This study explores these conditions through an agent-based model (ABM) that integrates the dynamics of attitude formation with the 'spiral of silence' theory. The simulations reveal that opinions can polarize or converge due to subtle contextual changes—such as changes in social connectivity or elite influence—even when the degree of attitude polarization remains moderate. Furthermore, the findings show that increased social connectivity attenuates the polarization of both attitudes and opinions, as greater exposure to diverse perspectives mitigates the effects of repulsion toward opposing views. These findings highlight how public opinions may fail to reliably reflect the true sentiments of the population, creating a misleading impression of a more fractured society while suggesting that increased connectivity could help mitigate such divisions.

Keywords

public opinion formation, spiral of silence, social networks

[1] Hanyang University, Seoul, Korea

Corresponding Author:
Dongyoung Sohn, Department of Media and Communication, Hanyang University, 222 Wangsimni-ro, Seongdong-gu, Seoul 133-791, Korea.
Email: dysohn@hanyang.ac.kr

The widespread adoption of digital media acts as a double-edged sword in communication, enhancing citizens' participation in public discourse while potentially amplifying polarization and populism (Lorenz-Spreen et al., 2022). Social media, in particular, raises concerns about deepening societal divides by fostering *echo-chambers*, where individuals primarily interact with like-minded others (Flaxman et al., 2016; Sunstein, 2017). Although diversifying the range of viewpoints presented to users has often been suggested as a remedy, recent studies indicate that encountering opposing opinions on social media may actually intensify polarization rather than mitigate it (Bail et al., 2018; Törnberg, 2022). Moreover, there is a concern that private attitudes may be more polarized than they appear, obscured by self-censorship or fear of social isolation (Hayes, 2007; Hayes et al., 2005; Mutz, 2002). This situation hints at a deeper societal fragmentation than what is immediately visible, as individuals hide their extreme viewpoints to avoid social consequences.

Conversely, an alternative perspective posits that the opinions prevalent on social media platforms may not accurately represent the variety of private beliefs, as evidence suggests that widespread attitude polarization is relatively uncommon (Baldassarri & Bearman, 2007; Prior, 2013). The concept of a 'silent majority', where individuals with moderate attitudes may opt for silence, leaving the public stage to vocal minority of extremists, exemplifies this point (Manfredi et al., 2020). Recent studies also indicate that exposure to polarized opinions can silence less extreme voices (Cinelli et al., 2021) and even discourage social media usage entirely (Nordbrandt, 2023), thereby effectively enlarging the volume of silence within the public sphere. This suggests that the appearance of highly polarized opinions might not accurately reflect the true sentiments of the population, which remain relatively invariant, because a substantial portion of the population may choose to conceal their true beliefs.

These differing perspectives highlight the complexity of the issue but converge on a crucial insight—the distributions of private attitudes and expressed opinions do not always match. This distinction, often overlooked in academic research, is essential for comprehending public opinion dynamics (Banisch & Olbrich, 2019; Manfredi et al., 2020). The visible landscape of opinions may significantly differ from actual attitudes, a notion tied closely with the *spiral of silence* theory (Noelle-Neumann, 1974), which explores how silence in the public sphere grows or recedes. However, much existing research on the spiral of silence has primarily focused on identifying the psychosocial factors leading to silence, often overlooking the intricate interplay between attitudes and publicly voiced opinions (Sohn, 2022; Sohn & Geidner, 2016).

The convergence or decoupling between underlying attitudes and expressed opinions largely depends on how and what kinds of *opinion climates* individuals face, which is in turn dependent on the composition of local social neighborhoods they belong to as well as the topology of global social networks. When individuals interact within echo-chambers of similar views on social media, they may mistakenly believe their attitudes are more widely shared than they actually are, known as the phenomenon of *false consensus* (L. Ross et al., 1977), and thus likely speak out them publicly. As illustrated by Mutz (2002), conversely, exposure to a diverse range of viewpoints might lead to *pluralistic ignorance*, where individuals incorrectly perceive their views as less common than they are (Katz & Allport, 1931). This misperception may lead to a reluctance to share their attitudes openly.

The spectrum of potential scenarios between these two distinct situations is vast, shaped by the intricate ways individuals form, modify, and choose to disclose their attitudes during interactions with others. This complexity suggests that the distributions of attitudes and opinions emerge not merely from a collection of individual states but as a result of iterative, complex interactions among numerous interconnected individuals (Galesic et al., 2021). Every individual in this extensive network influences and is influenced through a series of local connections, creating a web of influence that governs the evolution of attitudes and opinions. Given the scale and complexity of these networks, it can be extremely cumbersome, if not impossible, to track these dynamics over time with traditional methods, such as variable-based modeling (Macy & Willer, 2002; Miller & Page, 2007).

In seeking an alternative approach, this study employs an agent-based model (ABM) that integrates the dynamics of attitudes with the mechanisms of opinion expression, extending from the spiral of silence theory. The utilization of computer simulations has been proven to be particularly suited for studying the emergent processes involving complex social dynamics, such as public opinion and media effects (Cabrera et al., 2021; B. Ross et al., 2019; Sohn, 2022; Sohn & Geidner, 2016; Song & Boomgaarden, 2017; Waldherr, 2014). The primary objective of this study is to employ integrative simulations to explore under what conditions the distribution of publicly expressed opinions either polarizes or converges relative to the distribution of private attitudes within a networked communication environment.

Modeling the Interplay Between Attitudes and Opinions

Attitude is a private, evaluative response to an object or issue, reflecting an individual's internal beliefs or feelings. In contrast, an opinion is the outward expression of that attitude, manifesting as an observable statement or action, such as political arguments or votes. In public opinion research, these two terms are frequently used without clear distinction, and models that explore their dynamics typically adhere to one of two pathways: *comparison-based processes*, where individuals adjust their attitudes/opinions by evaluating those of others, and *alignment-based processes*, where attitudes/opinions shift to align with the majority or minority. Comparison-based processes, which are presumed to occur at the inter-individual level, can be modeled in various ways—using either discrete states, like in the Ising model (Li et al., 2019), or continuous variables that change in pairwise interactions (Deffuant et al., 2000) or group dynamics (Hegselmann & Krause, 2002). The continuous variable approach often features the concept of *bounded confidence*, depicting individuals' tendency to align with similar opinions and ignore those that significantly deviate from their existing beliefs (Deffuant et al., 2000; for comprehensive review of these models, see Castellano et al. (2009).

However, empirical evidences suggest that individuals might not only disregard views significantly different from their own but could also actively adjust their attitudes to further distance themselves from conflicting perspectives (Bail et al., 2018; Myers & Bishop, 1970; Sherif & Hovland, 1961). Building upon these insights, several models have been proposed that encapsulates the dual processes of assimilation and differentiation (Keijzer et al., 2024). For instance, Macy et al. (2003) adapted the Hopfield network model to simulate how individuals are drawn toward similar others and repelled by dissimilar ones, providing an early framework for the dual processes. Flache and Macy (2011) further expanded this approach to investigate how opinions within social networks assimilate or diverge, particularly in the context of long-range social connections. More recently, Axelrod et al. (2021) applied the attraction-repulsion rule to exploring the various circumstances that lead to the polarization or convergence of attitudes.

In models grounded in alignment-based processes, meanwhile, attitude adjustments occur between an individual and a group, depending on alignment with either majority, or minority group norms. *Social Impact Theory* illustrates this, emphasizing that attitudes are shaped by the strength, immediacy, and number of sources of social influence (Latané & Wolf, 1981; Nowak et al., 1990). These principles suggest that in majority contexts, individual attitudes are reinforced through the cumulative effect of multiple sources of social influence. Conversely, in minority settings, the relative lack of these reinforcing influences can lead to a weakening of these attitudes, a process akin to the spiral of silence (Noelle-Neumann, 1993). Leveraging these insights, a range of models have been developed, including discrete state models like the voter model (Clifford & Sudbury, 1973) and the majority rule model (Galam, 1997) and continuous variable models, to explore phenomena such as the spiral of silence (B. Ross et al., 2019; Sohn, 2022; Sohn & Geidner, 2016) and selective exposure to media (Song & Boomgaarden, 2017).

While these models undoubtedly offer valuable insights, they frequently conflate the concepts of attitudes and opinions, overlooking the intertwined relationship between them. Failing to discern these elements can yield an incomplete understanding of social dynamics, potentially leading to ineffective decisions or misinterpretations (Manfredi et al., 2020). For instance, businesses that consider only public customer reviews may fail to account for more enduring, private attitudes shaped by factors like brand loyalty, potentially leading them to overprioritize short-term gains rather than building long-term brand value. Likewise in politics, focusing on the loudest opinions can push positions to extremes, neglecting moderate or ambivalent attitudes that remain publicly unspoken. Conversely, attending only to majority attitudes risks overlooking the crucial insights of vocal minorities (Moscovici, 1976; Prislin, 2022). As these examples illustrate, conflating private attitudes with publicly expressed opinions may result in misattributing silence to broad consensus or polarization, when in reality many hold unvoiced, divergent beliefs. The need to bridge this gap is evident—attitudes and opinions are inseparably linked, and understanding this relationship is pivotal for comprehending the various emergent patterns in social contexts, particularly in platforms like social media where a substantial number of individuals choose silence (Sohn & Choi, 2023).

Moreover, the conflation of attitudes and opinions in these models presents a false dichotomy between comparison-based and alignment-based processes as mutually exclusive mechanisms. This overlooks the potential for these processes to coexist in shaping attitudes and opinions, with their occurrence varying depending on social contexts (Kendal et al., 2018). Alignment-based processes, for example, become particularly relevant under group pressures where individuals might mask their true attitudes (Cialdini & Goldstein, 2004; Sassenberg & Jonas, 2007), suggesting its utility in modeling opinion expression rather than attitude adjustment, a domain where comparison-based processes might be more applicable. Distinguishing between private attitudes and expressed opinions, therefore, reveals two distinct processes—adjusting attitudes, which involves internal belief changes, and deciding to express opinions, which concerns the public articulation of these beliefs. Each requires its own mechanism, comparison for the former and alignment for the latter, highlighting the need for models that accommodate both aspects to fully capture the dynamics of social influence.

The Roles of Social Networks

Understanding the transition from private attitudes to expressed opinions requires an exploration of not only how attitudes are formed and changed but also the contexts and conditions under which individuals choose to reveal or conceal their attitudes (Cialdini & Goldstein, 2004; Noelle-Neumann, 1993). Research into the spiral of silence has extensively explored the reasons why individuals choose to stay silent when confronted with majority opinions. Several moderating factors have been identified, including self-censorship (Hayes, 2007), communication apprehension (Neuwirth et al., 2007), attitude certainty (Matthes et al., 2010), disagreement and publicness (Chen, 2018), among many others (for metanalytic review, see Matthes et al. (2018)).

In addition to these psychological moderators, the decision to express opinions also hinges on the size and structures of the surrounding social networks (Sohn, 2022). When people have greater social reach—the spatial and structural range within which they form ties with others—they are more likely to encounter diverse opinion climates where majority pressures tend to be diluted. Since larger networks are more likely to include weak ties (Granovetter, 1973), the perceived social pressure from the majority may be reduced when those individuals are not closely connected. Conversely, limited social reach may confine individuals to more homogeneous environments, reinforcing dominant views and discouraging dissent. Overall, these patterns suggest that individuals' willingness to publicly express their attitudes is shaped in part by their social reach and the structural composition of their networks.

Additionally, individuals who are centrally located in a network, often termed 'opinion leaders' or simply 'elites', can yield greater influence over attitude change and opinion climates due to their numerous direct or indirect connections (Centola, 2010; Watts & Dodds, 2007). Many studies have already shed light on the roles of social network structures and elites in shaping public opinion (Watts & Dodds, 2007), the behavioral diffusion process (Assenova, 2018; Centola, 2018), the alignment of political issue domains (Baldassarri & Bearman, 2007), and the spiral of silence process (Cabrera et al., 2021; Sohn, 2022; Sohn & Geidner, 2016). However, beyond their mere presence, the distribution of their voices—their level of consensus or disagreement on specific issues—can significantly impact the formation of public opinion. For instance, recent studies found that unless there was convergence in news media opinions, their impact quickly dissipated in the spiral of silence process (Sohn, 2022). This phenomenon can be partially explained by the concept of *information entropy*, a measure of uncertainty or disorder in an environment (Shannon, 1948). The mere presence of alternative viewpoints in a discourse can significantly increase the level of information entropy, introducing a greater degree of uncertainty into the opinion landscape.

Besides the consensus or lack thereof, the intensity and extremity of the opinions these elites express can significantly determine the direction of public sentiment. For instance, extreme elites—those who hold and express views at the far ends of an attitude spectrum—may exacerbate affective polarization (Druckman et al., 2021). They can attract individuals with similar attitudes while repelling those with differing attitudes (Axelrod et al., 2021), which is particularly pronounced among those holding strong partisan attitudes or less politically informed (Bäck et al., 2023; Banda & Cluverius, 2018). The presence of extreme elites could also embolden individuals whose attitudes are in the minority to voice their opinions publicly. As a result, these extreme elites can make the overall opinion climate appear more polarized, potentially impeding the convergence of individual opinions (i.e., the spiral of silence) and leading to the perception of a more divided public.

In contrast, the presence of moderate or ambivalent elites—those with viewpoints near the middle of the attitude spectrum—could potentially mitigate the tendency of individuals to express extreme viewpoints (Röchert et al., 2022). Just as experts can mitigate the diffusion of misinformation through corrective interventions (Lewandowsky et al., 2017), or framing polarization as problematic can prompt people to support bipartisanship (Robison & Mullinix, 2016), moderate elites may similarly exert a moderating influence on public discourse. In a networked communication environment, these extreme or moderate elites might serve as guideposts toward which individual attitudes and opinions are attracted or repelled (Axelrod et al., 2021). Extreme elites could pull ambivalent attitudes toward more extreme positions, while moderate elites could potentially attenuate affective polarization by allowing space for diverse perspectives.

Drawing on the issues discussed above, this study aims at exploring the following research questions:

RQ1. Under what conditions do the distributions of private attitudes and expressed opinions converge or diverge, and when they diverge, which tends to show greater polarization?

RQ2. Does increasing social reach—thereby increasing the average degree of individual networks—mitigate or exacerbate the polarization of attitudes and opinions?

RQ3. What role do the diversity and extremity of elite opinions play in the dynamics of attitudes and opinions?

In the following sections, a model of attitude formation and opinion expression and simulation procedures are described in detail.

Model Development

The evolution of public opinion results from iterative interactions among a multitude of individual actors holding private attitudes dispersed across social networks. Therefore, to examine how these dynamics unfold, we need to adopt a bottom-up approach that initially establishes the rules for individual behaviors, specifically regarding the formation and change of attitudes through interactions with others (Miller & Page, 2007). Consider a society of N individuals, each holding an attitude toward a contentious issue, such as the mandatory vaccination or the legalization of same-sex marriage. Some might support it, while others may oppose it, all with varying levels of intensity. Building upon the work of Sohn (2022), the attitude of an individual i (denoted by a_i) is modeled as a composite of two elements: valence and confidence. Valence is a dichotomous variable with either a $+1$ or -1 value, signaling the direction of the attitude, while confidence is a continuous vector with real values in [0, 1] range. When combined, these factors allow for the attitude value to range in [−1, 1].

Attraction-Repulsion Rule for Attitude Change

Attitudes are not formed in isolation but are subject to various social influences (Cialdini & Goldstein, 2004; Sassenberg & Jonas, 2007). When people come across opinions that match their pre-existing attitudes, they often adjust their attitudes to align more closely with those views. Conversely, when faced with opinions that greatly differ from their own, they may shift their attitudes to further distance themselves from those conflicting views (Keijzer et al., 2024; Macy et al., 2003). To contextualize the dual processes, it is necessary to assume that individuals possess a psychological tolerance or latitude that establishes the limit for tolerating dissimilar opinions—below this limit, individuals assimilate with others, while beyond it, they distance themselves. Suppose we can quantify a person's tolerance latitude on an attitudinal scale, represented by ϵ, which indicates the maximum extent of difference they can tolerate. For instance, an individual might have a tolerance level set at 1.0 on an attitudinal scale spanning from -1 to $+1$. This means that he or she can bear any opinion from others, as long as the distance from their own attitude is equal to or less than 1.0.

To put this formally, let's consider the attitude of an individual i at time $t+1$ is a function of their prior attitude at time t, and the absolute difference between their own attitude and the opinion of a randomly selected social neighbor j, denoted as $d_{ij}^{(t)} = \left| a_i^{(t)} - o_j^{(t)} \right|$. Here, $d_{ij}^{(t)}$ represents the distance in attitudes between two individuals on the attitudinal scale only if individual j 's attitude is outwardly revealed; $o_j^{(t)}$

represents individual j 's expressed opinion at time t. If $d_{ij}^{(t)}$ is equal to or smaller than ϵ, suggesting that individual i finds the neighbor j 's opinion within their tolerance limit, then individual i's attitude at time $t+1$ may shift closer toward the arithmetic mean of their current attitudes, $\mu_{ij}^{(t)}$. Conversely, if $d_{ij}^{(t)}$ exceeds ϵ, indicating the difference is beyond i 's tolerance, individual i 's attitude at $t+1$ will diverge from j 's opinion by an amount proportional to the difference between their mean and i 's current position.

A limitation of the described attitude adjustment model is that large differences in attitudes can lead to disproportionately large shifts, which contradicts empirical findings suggesting that attitudes—especially extreme ones—are generally more resistant to change (Tormala & Petty, 2004). To reflect the diminishing likelihood of significant changes as attitudes approach the extremes, it is necessary to incorporate a responsiveness parameter, $\gamma\left(a_i^{(t)}\right)=\frac{s}{1+\left|a_i^{(t)}\right|}$, which reduces the magnitude of both attraction and repulsion in attitude adjustments as attitudes become more extreme. Here, $s=0.5$ sets the maximum responsiveness when attitude is near zero, gradually decreasing as the absolute value of the attitude increases. Additionally, to ensure attitudes remain within the predetermined range of [−1, 1], a damping factor, $\kappa\left(a_i^{(t)}\right)=1-\left|a_i^{(t)}\right|^{\beta}$, is applied in situations where attitudes are further apart near the boundaries, with the parameter β set at 1.0. The rule can be summarized as follows (see Appendix for more details):

$$a_i^{(t+1)}=\begin{cases} a_i^{(t)}+\gamma\left(a_i^{(t)}\right)\left(\mu_{ij}^{(t)}-a_i^{(t)}\right), if\ d_{ij}^{(t)}\leq\epsilon \\ a_i^{(t)}-\gamma\left(a_i^{(t)}\right)\left(\mu_{ij}^{(t)}-a_i^{(t)}\right)\kappa\left(a_i^{(t)}\right), if\ d_{ij}^{(t)}>\epsilon \end{cases} \tag{1}$$

Opinion Expression Rule

Not all attitudes are outwardly expressed, with only a fraction made publicly observable as opinions. In modeling this process, especially in the spiral of silence literature, it has been a common assumption that individuals maintain fixed thresholds for expression and only reveal those attitudes that surpass them (Cabrera et al., 2021; B. Ross et al., 2019; Sohn, 2022; Sohn & Geidner, 2016). To formally define the threshold rule, let an individual i 's expression threshold be denoted as ϕ_i, which varies across individuals but remains static over time. The probability of this individual revealing their own attitude can then be expressed as follows:

$$P(expression)_i^{(t)}=\begin{cases} 1\ if\left|a_i^{(t)}\right|>\phi_i \\ 0\ if\left|a_i^{(t)}\right|\leq\phi_i \end{cases} \tag{2}$$

Individual attitudes, typically modeled as comprising valence (direction) and confidence (strength) on a scale, become more extreme as confidence increases in either direction. This equation formalizes that the likelihood of an individual publicly revealing their attitude depends solely on the degree of confidence, irrespective of its direction. While it is reasonable to assume that individuals reveal their attitudes only when confident enough, previous studies have typically assumed that this confidence changes based on individuals' majority or minority status—introducing a separate mechanism for attitude change that diverges from the attraction-repulsion model described earlier. This presents a conceptual inconsistency, as it involves two distinct mechanisms to explain the same process of attitude change.

To integrate these two different mechanisms in a single model, we propose to change the once-static threshold for expression into an adaptive variable, $\phi_i^{(t)}$, that adjusts over time in response to majority-minority dynamics an individual encounters. This modification can be implemented with a slight adjustment to the logistic function used in prior models (B. Ross et al., 2019; Sohn, 2022; Sohn & Geidner, 2016), as indicated in Equation 3.

$$\phi_i^{(t)} = \frac{l}{1 + e^{\tau\delta_i^{(t)}}} \tag{3}$$

Here, $\phi_i^{(t)}$ represents individual i 's expression threshold at time t, which dynamically shifts in response to the surrounding opinion climate. This climate is quantified by the local opinion ratio, defined as: $\delta_i^{(t)} = \frac{n_s^{(t)} - n_o^{(t)}}{n_s^{(t)} + n_o^{(t)}}$, where $n_s^{(t)}$ and $n_o^{(t)}$ represent the counts of directly observable opinions that share the same or opposite direction (i.e., $+1$ or -1) as individual i 's own stance, irrespective of their confidence or strength.[1] Consequently, $\phi_i^{(t)}$ rises in minority scenarios (i.e., $\delta_i^{(t)} < 0$), reflecting a hightened reluctance to reveal one's own attitude, but falls in majority ones (i.e., $\delta_i^{(t)} > 0$), indicating a higher propensity to share one's opinion.

In this integrated model, there are thus two central variables—individuals adjust their attitudes by comparing them with their randomly selected social neighbors while concurrently modifying their expression thresholds according to whether their attitudes align with the local majority or minority opinions. This method allows us to holistically examine how private attitudes evolve through social comparison processes and how these private attitudes transition into publicly expressed opinions as dictated by the principles of the spiral of silence theory. For example, if an individual perceives attitude being in the minority (or majority), the threshold for expressing it increases (or decreases), rather than the attitude itself changing. This means that even when an individual feels strongly about a particular issue, they may choose to keep it to themselves if they perceive themselves to be in the minority. This phenomenon, which is frequently observed in real-world situations, was not sufficiently represented in the earlier models.

Polarization Indices for Private Attitudes and Expressed Opinions

Numerous formal measures to quantify polarization have been discussed in the literature, with Bramson et al. (2016) providing a comprehensive review. In line with our simulation's primary goal of exploring the convergence or decoupling between attitude and opinion distributions, a composite index is proposed to assess the degree of polarization in both private attitudes and publicly expressed opinions. This index, the attitude polarization index (API), is calculated using the product of the attitude variance at time t, denoted as $\sigma_{att}^{(t)}$, and the inverse of coverage, $1-\lambda_{att}^{(t)}$. The index increases as variance grows and decreases as coverage increases.

$$API = \sigma_{att}^{(t)}\left(1-\lambda_{att}^{(t)}\right) \tag{4}$$

The term, 'coverage', refers to the proportion of the attitudinal range that is actively occupied by individuals' attitudes. It measures the extent to which attitudes are distributed across the entire attitudinal scale, as opposed to being clustered in specific regions (Bramson et al., 2016). Coverage is calculated by dividing the number of distinct attitude bins occupied by the total number of possible bins in the attitudinal range. A high coverage value indicates that attitudes are widely distributed across the entire scale, reflecting greater diversity, while a low value suggests that attitudes are concentrated in specific areas, leaving other parts unoccupied.[2]

One advantage of the API is its ability to distinguish true polarization from other distributions that exhibit comparable spread. Although extreme clustering can yield marginally higher variance than a uniform distribution, this difference is often too subtle to be analytically useful. By weighting variance by the inverse of coverage (i.e., 1—coverage), the API amplifies signals of polarization when attitudes are both widely spread and concentrated at the extremes, leaving the middle sparsely populated. This allows the API to more accurately capture polarization and differentiate it from mere attitudinal diversity. Conversely, consensus or convergence emerges when both variance and coverage are low, while a distribution with both high variance and high coverage reflects broad but non-polarized diversity (p. 15).[3]

For the index of opinion polarization (OPI), the same API formula will be applied, but only to those attitudes that are publicly expressed—that is, attitudes that exceed individuals' expression thresholds. By restricting the calculation of API to publicly expressed attitudes, the index quantifies the polarization of opinion distributions based on attitudes that manifest overtly in public settings. This approach allows for consistent quantification of both attitude and opinion polarization, simplifying comparisons and enabling a clearer understanding of the circumstances under which private attitudes and publicly expressed opinions converge or diverge.

Simulation Settings and Procedures

This simulation begins by establishing a society of individual agents (N = 1,000) distributed over a two-dimensional toroidal plane, constructed with a simulation program, *NetLogo* version 6.3.0 (Wilensky, 1999).[4] There are global parameters that can be varied for experimentation, with the first being *social reach*, defined as a circular area with a specified radius. Each individual is initially assigned a social reach, and connections are formed exclusively between individuals located within overlapping circular areas, ensuring that both parties are within each other's social reach and can reciprocate the connection (Hamill & Gilbert, 2009, 2010). Varying the social reach parameter affects both the average size of individuals' local networks and the structural properties of the global network as a whole (see Table 1). This procedure generates networks that follow a slightly right skewed distribution of degrees, known to approximate the distribution of connection in ordinary social networks such as friendships or workplace relationships (Broido & Clauset, 2019; Newman et al., 2001; Rolfe, 2014; Sohn, 2022).[5]

The second parameter is the level of tolerance ϵ, which ranges between 0 and 2.0, meaning that once the level of ϵ is determined, everyone in the simulation has the same tolerance level. With $\epsilon = 1.5$, for example, a person with an attitude score of 1.0, an extreme positive attitude, sees a neighbor whose opinion score is -0.4 as acceptable as the distance is smaller than ϵ (i.e., $d_{ij} = 1.4 < \epsilon$). With $\epsilon = 0.5$, meanwhile, the same person with an attitude score of 1.0 finds a neighbor whose opinion score is 0.3 as intolerable (i.e., $d_{ij} = 0.7 > \epsilon$). Given that the maximum distance in attitudes is capped at 2.0, a higher tolerance level thus leads to individuals being more likely to encounter attractive opinions of others whereas a lower tolerance level increases the likelihood of individuals encountering opinions that they find repulsive.

In the simulation, another adjustable parameter is the presence of prominent individuals, termed 'elites', characterized by their greater social reach and influence compared to others. When elites are present, a random sample of 1% of the entire population (10 out of 1,000 agents) is selected and assigned a social reach that is 1.5 times greater than the rest of the population. These elites have fixed attitudes that are always publicly revealed, making them immune to both the attraction-repulsion rule and majority-minority dynamics. The distribution of elite opinions can be varied as well in three different ways—they may unanimously embrace one position ($10:0$), the majority supports one position while there is a minority position (e.g., $7:3$), or even split ($5:5$). Lastly, we can modify the extremity or intensity of elite opinion. In the extreme condition, all elites' opinions are set to the maximum values, either $+1$ or, while in the moderate condition, their opinions are diluted to the values closer to the middle (e.g., -0.25 or $+0.25$).

At the start of the simulation, agents are assigned random initial attitudes, $a_i \sim U(-1,1)$, and expression thresholds, $\phi_i \sim U(0,1)$, both drawn from a uniform distribution. Each agent then randomly selects a neighbor within their local network and evaluates whether the difference between their own attitude and the neighbor's expressed opinion falls within their tolerance threshold ϵ. This evaluation only takes place if the neighbor has publicly expressed their attitude; silent neighbors are excluded from consideration. Agents express their attitudes publicly if their attitude surpasses their expression threshold, making them observable to others, while they opt for

Table 1. Social Reach and the Structural Properties of Networks Generated.

Social reach	Average degree-non opinion leaders (std. dev.)	Average degree-opinion leaders (std. dev.)	Network density (%; std. dev.)	Global clustering (%; std. dev.)	Average path-length (std. dev.)	Network assortativity
15	6.90 (0.04)	15.77 (0.00)	0.35 (0.00)	58.17 (0.36)	∞	0.52
20	12.24 (0.08)	27.81 (0.00)	0.62 (0.00)	58.27 (0.37)	8.13 (0.22)	0.47
25	19.12 (0.12)	43.45 (0.00)	0.96 (0.00)	58.27 (0.37)	6.13 (0.19)	0.44
30	27.51 (0.17)	62.17 (0.00)	1.39 (0.00)	58.31 (0.36)	4.98 (0.17)	0.41

Note. The reported figures are from 20 replications at each level of social reach. The average path-length of networks with a social reach of 15 is considered infinite, as these networks consist of several disconnected components.

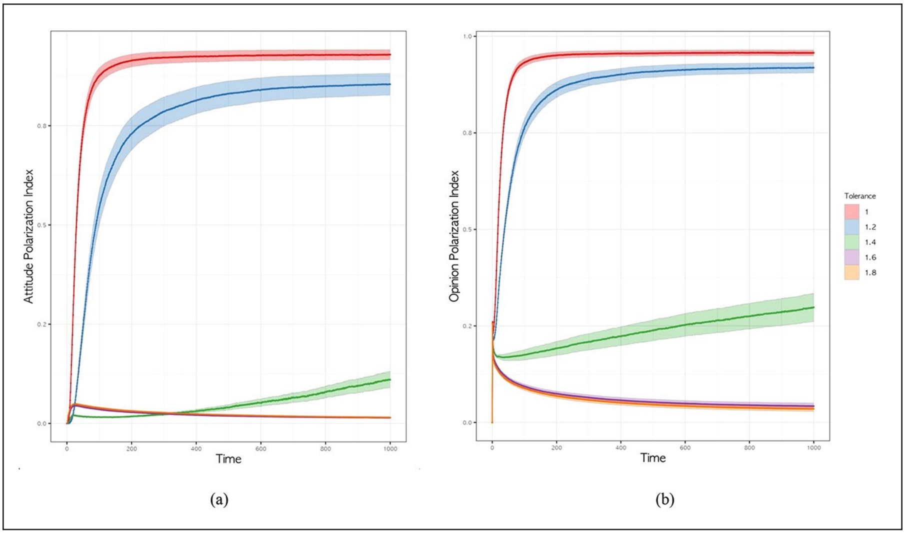
silence otherwise. This procedure was repeated 1,001 times for each simulation run (from $t = 0$ to $t = 1,000$). Each simulation run was then replicated 20 times for each of the 288 global parameter combinations, which were derived from the following factors—6 levels of tolerance, 4 levels of social reach, 2 conditions of elite presence, 3 configurations of elite-opinion distribution, and 2 levels of elite opinion strength. Altogether, this generated a total of 5,765,760 data points.[6]

Results

To delve deeper into the effects of varying parameters (see Appendix for sensitivity analyses results), it is helpful to examine the simulation outcomes visually. Figure 1 illustrates the changes in polarization indices for attitudes and opinions over time across different tolerance levels, with the data averaged over 20 simulation replications for each tolerance level. At lower tolerance ($\epsilon \leq 1.2$), both measures peak, indicating the state of polarization, while at higher tolerance ($\epsilon \geq 1.6$), they decrease and stabilize, showing convergence. This generally aligns with prior findings suggesting that attitudes can quickly converge or diverge outside of certain tolerance band (Axelrod et al., 2021). In between these cases, there exists a vast valley—at a mid-range tolerance ($\epsilon = 1.4$), both API and OPI remain relatively stable over time, increasing only gradually and leveling off at approximately 0.1 and 0.2, respectively, by the end. Notably, the API remains consistently lower than the OPI over time, particularly at a tolerance level of 1.4, suggesting that expressed opinions appear more diverse or widespread than the underlying attitudes.

This observation is further illustrated in Figure 2a, which compares both polarization indices and the proportion of silence at the end of the simulations (t = 1,000) across different tolerance levels. The OPI consistently exceeds the API, with the gap becoming most pronounced around a tolerance level of $\epsilon = 1.4$, coinciding with a dramatic increase in the proportion of silent population. Under conditions of severe polarization at lower tolerance levels, attitudes and opinions tend to converge closely, reducing the likelihood of a spiral of silence. However, as the tolerance level increases, phase transitions emerge around $\epsilon = 1.4$, where opinions and attitudes diverge significantly. At this point, more individuals opt for silence rather than expressing their attitudes. If this silence is disproportionately adopted by those with moderate (extreme) attitudes, public discourse may appear more (less) polarized than the underlying distribution of attitudes. A natural question that follows is whether those choosing silence are primarily individuals with moderate or extreme attitudes, a factor critical in shaping the visible opinion landscape.

Figure 2b presents the individual-level distributions of attitudes and expressed opinions at the end of the simulations across three different tolerance levels (excluding those below $\epsilon = 1.4$, as both attitudes and opinions are highly polarized). Expressed opinions (blue) are more widespread than attitudes overall (red), consistent with the observation in Figure 2a that the OPI consistently exceeds the API. Interestingly, the distribution of silent attitudes (green) is more spread out than the overall attitude distribution, suggesting that silence encompasses not only ambivalent or moderate views



(a)

(b)

Figure 1. Temporal change of attitude and opinion polarization across different tolerance levels.

Note. The figure illustrates the evolution of (a) the attitude polarization index (API) and (b) the opinion polarization index (OPI) over time across varying tolerance levels. Lower tolerances ($\epsilon \leq 1.2$) result in rapid polarization, whereas higher tolerance ($\epsilon \geq 1.6$) promotes convergence. At an intermediate tolerance level ($\epsilon = 1.4$), the OPI stabilizes at approximately 0.2, while the API remains low initially, but gradually increases to around 1.0.

but also strongly held extreme ones. Rather than reflecting a lack of conviction, this pattern often arises because individuals perceive themselves to be in a local minority, which raises their expression thresholds, and leads them to withhold even resolute positions. Taken together, these findings align with the perspective that underlying attitudes are generally less polarized than the public opinions that surface—not simply because moderates refrain from speaking, but also despite the fact that some highly polarized individuals choose not to voice their views when they feel outnumbered (a phenomenon often referred to as ‘loud silence’).

To further explore the contextual conditions for phase transitions, Figure 3 illustrates how API and OPI vary with changes in two key variables: social reach and the presence of elites, while keeping the tolerance level fixed at $\epsilon = 1.4$. In Figure 3a, API decreases as social reach—which determines the size of individuals' networks—increases, especially when elites are absent (shown in blue). This depolarizing effect is more evident in Figure 3b, where OPI shows a sharp decline as network sizes expand. These patterns reflect changes in the opinion climates surrounding individuals—as networks grow, the surrounding opinion climates are likely to become more diverse and heterogeneous, as the probability of encountering entirely homogeneous opinions decreases rapidly. This diversification may elevate individuals' expression thresholds, reducing the visibility of extreme opinions, which, in turn, fosters convergence in private attitudes and ultimately decreases polarization.

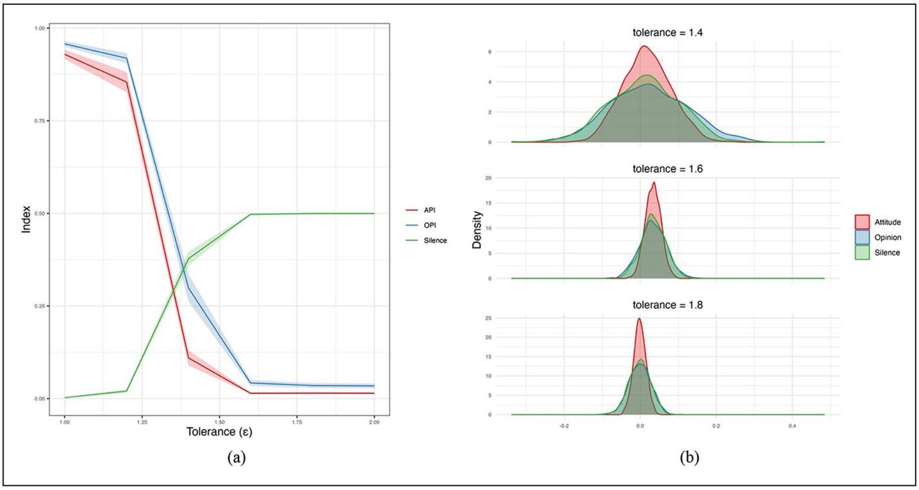
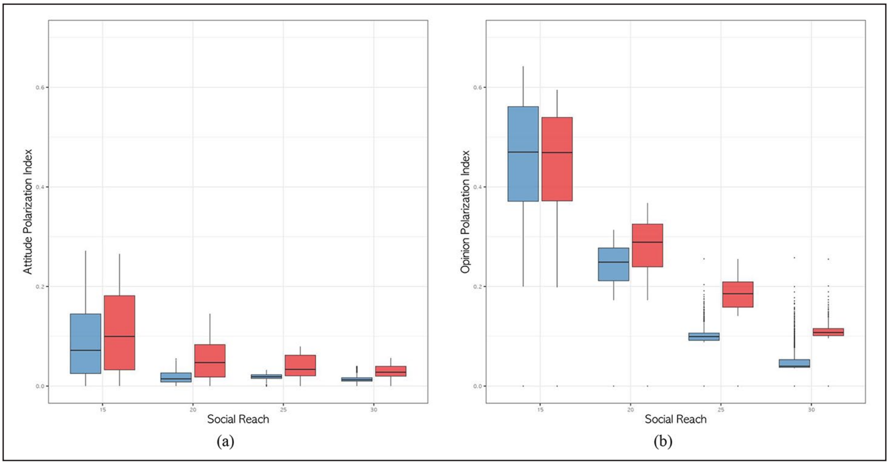


Figure 2. Polarization indices and attitude distributions across different tolerance levels.
Note: Panel (a) shows the attitude polarization index (API), opinion polarization index (OPI), and the proportion of the silent population across varying tolerance levels (ϵ) at time = 1,000. Panel (b) illustrates the distributions of private attitudes, expressed opinions, and the attitudes of silent individuals at three representative tolerance levels (ϵ = 1.4, ϵ = 1.6, and ϵ = 1.8). These agent-level attitude distributions were captured at time = 1,000 and averaged over 20 repetitions. For clarity, the distributions corresponding to a tolerance level of ϵ < 1.4 have been omitted from part (b).

When elites are introduced (in red), however, the declines in both API and OPI become more gradual, suggesting that elites may counteract the depolarizing effects of increased social connectivity. That is, the presence of elites, whether moderate, or extreme, can disrupt the convergence of attitudes and opinions and potentially inhibit the spiral of silence effect. By attracting or repelling individuals—including those with initially neutral or ambiguous attitudes—toward more polarized positions, elites may homogenize rather than diversify opinion climates surrounding individuals. This, in turn, could stimulate greater vocal participation and reduce the prevalence of silence. However, does this mean that the presence of elites universally obstruct the convergence of attitudes and opinions, and entirely disrupts the spiral of silence process? This proposition appears to conflict with prior research, which has shown that the news media, a form of elite, can actually facilitate the spiral of silence, especially when they present a unified or homogeneous view (Slater, 2007; Sohn, 2022).

To investigate this issue, the variations in API and OPI were analyzed in relation to the diversity and extremity of elite opinions. Figure 4 illustrates these changes under three scenarios of elite opinion diversity—unanimous (10:0), uneven (7:3), and evenly split (5:5)—and contrasts between moderate (in blue) and extreme elites (in red). Figure 4a reveals that the presence of extreme elites significantly broadens the range of possible attitude polarization, in contrast to the condition with moderate elites, where API stays minimal. This pattern remains consistent regardless of the diversity of elite opinions but becomes less pronounced as social connectivity increases. Figure 4b, similarly,



(a)

(b)

Figure 3. Impact of elites on polarization across different social reach levels.
Note. The tolerance is fixed at $\epsilon = 1.4$. The figures showcase boxplots that reveal how the presence of elites (in red) versus their absence (in blue) affects (a) attitude polarization and (b) opinion polarization at varying levels of social reach.

showcases a marked reduction in OPI with the expansion of network sizes, a trend particularly pronounced in the presence of moderate elites. Conversely, extreme elites appear to mitigate the depolarizing influence that might otherwise result from greater social connectivity. These findings indicate that the influence of elites on the polarization of attitudes and opinions hinges on the extremity of their stances and the breadth of individual networks. Specifically, moderate elites may facilitate the convergence of attitudes, reducing the visibility of extreme opinions—an effect that is amplified when individual networks enlarge. Conversely, extreme elites intensify attitude polarization, increasing the visibility of extreme attitudes and counteracting the converging influence of enhanced social connectivity.

Figure 5 features contour plots that delve into the circumstances leading to convergence or decoupling between attitudes and opinions, quantified as OPI minus API, examining a spectrum of the levels of tolerance and social reach. These differences are visually represented, with greater differences highlighted in shades of red, and smaller disparities depicted in shades of blue. Figure 5a illustrates conditions without elites, showing that larger discrepancies between attitudes and opinions, indicated by red shades, tend to emerge around tolerance levels of 1.2 and 1.4, especially within smaller networks. This suggests that when low-tolerance individuals happen to be in compact networks, the gap between API and OPI widens, making expressed opinions appear more polarized than underlying attitudes. This effect may result from limited exposure to diverse opinions in smaller networks, where individuals are more likely to encounter one-sided opinion climates. This environment not only facilitates attitude convergence, but also lowers expression thresholds for majority-aligned individuals, leading to a dispersed but skewed distribution of expressed opinions. As tolerance increases beyond 1.4 and social connectivity improves, the API-OPI gap diminishes.

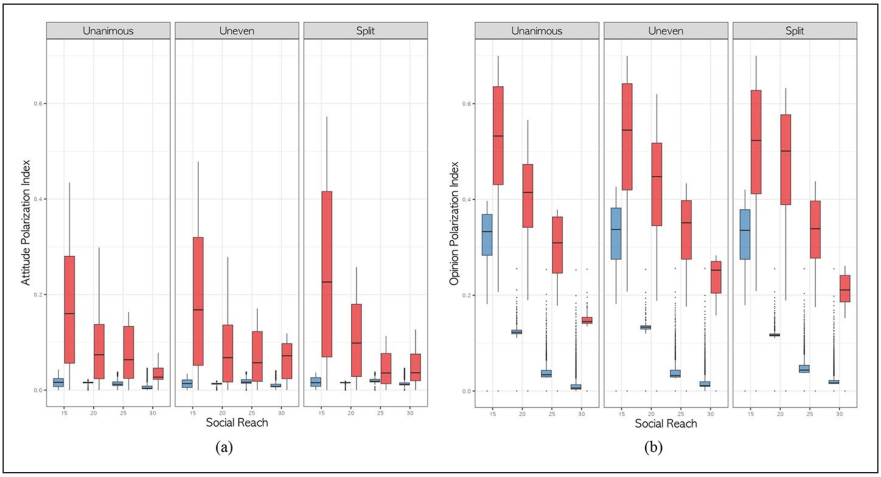


Figure 4. Influence of extremity and diversity of elites on polarization across different levels of social reach.

Note: With the tolerance level set at $\epsilon = 1.4$, this figure depicts how (a) attitude polarization and (b) opinion polarization are influenced by the presence of elites with extreme (red) versus moderate (blue) viewpoints, across different distributions of opinions (unanimous/uneven/split).

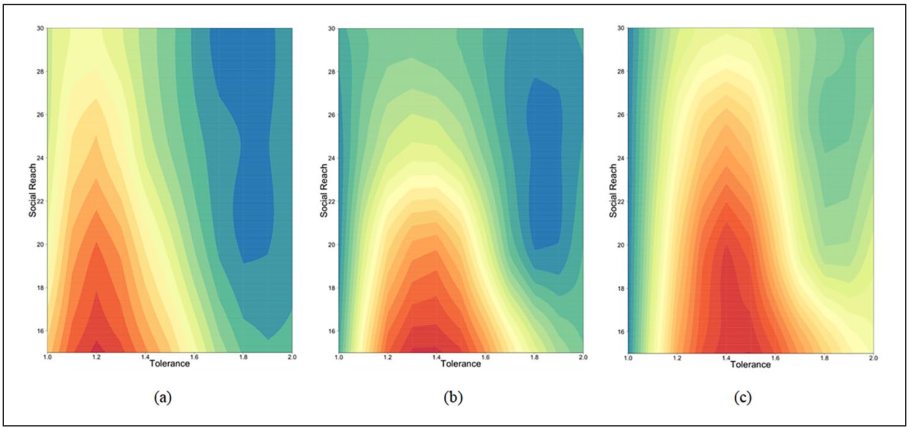


Figure 5. Contour plots of the disparities between attitude and opinion polarization.

Note. These contour plots illustrate how the difference between the polarization indices of attitude and opinion (i.e., OPI-API) vary as functions of tolerance and social reach, (a) without the presence of elites, with the presence of (b) moderate elites and (c) extreme elites. Areas of higher values are represented in red, while those with lower values are depicted in blue.

In Figure 5b, the introduction of moderate elites slightly shifts the red areas to cover a broader range of tolerance levels from 1.2 to 1.6, particularly in smaller networks. This widening may result from moderate elites promoting further attitude convergence, while smaller networks, in parallel, lower expression thresholds for majority-aligned individuals, promoting the expression of more extreme opinions. These opposing dynamics—attitude convergence and the polarization of expressed opinions—amplify the observed disparities between attitudes and opinions. Figure 5c illustrates that, under the influence of extreme elites, the regions illustrating discrepancies expand considerably, dominating much of the plot. This indicates a greater likelihood of opinions diverging substantially from underlying attitudes, even within larger networks, except at extremely low tolerance levels. This phenomenon may arise because extreme elites attract or repel individual attitudes in opposing directions. As network size increases, expression thresholds also rise, leading most individuals to refrain from expressing their views unless their attitudes are strongly held. Consequently, only highly polarized attitudes are publicly expressed, resulting in a more extreme distribution of expressed opinions than of underlying attitudes, thereby expanding the regions of attitude-opinion decoupling.

Discussion and Implications

Attitudes, being inherently private, are not directly observable by others. We infer them primarily through the prism of expressed opinions, the outward manifestation of inner beliefs. The current simulation results demonstrate that, as attitudes are selectively revealed, the opinion landscape may often diverge from the actual distributions of individual beliefs, a finding consistent with prior modeling work (Banisch & Olbrich, 2019). Specifically, the findings indicate that opinion distributions are generally more dispersed or polarized than the underlying attitudes, aligning with one of the competing perspectives discussed earlier (Baldassarri & Bearman, 2007). This pattern reflects a phenomenon often referred to as ‘polarization without persuasion’, where overt behaviors, such as expressed opinions, or voting, appear more polarized or widespread than underlying attitudes—a pattern documented in many empirical studies (see Prior, 2013, for review).

A particularly relevant example can be seen in the case of COVID-19 vaccination in the United States. While public discourse in the media often reflected deep polarization, polling data indicate that most individuals did not hold extreme pro- or anti-vaccination views. In fact, over 60% of Americans reported that the benefits of COVID-19 vaccination outweighed the risks while exhibiting vaccine hesitancy or cautious support, often stemming from concerns about potential side effects or mistrust in medical authorities or government institutions (Funk et al., 2023). This decoupling between private attitudes and public expressions may be more widespread than commonly assumed. The current simulations help clarify how such divergence can occur by illustrating a potential mechanism through which expressed opinions become polarized, even in the absence of corresponding attitude polarization.

The results also indicate the presence of a critical tolerance band (i.e., $\epsilon \approx 1.4$), where uncertainty in the direction of attitude shifts increases, leading to more complex and less predictable dynamics between attitudes and opinions. Factors such as the presence and extremity of elite voices or changes in social connectivity can serve as tipping points, effectively breaking the previously observed coupling between attitudes and expressed opinions. Analogous to how water solidifies around the freezing point of $0°C$, this suggests that a critical tolerance level sets the stage for dynamic phase transitions in response to subtle contextual variables. While it remains unclear what the critical level of tolerance represents in reality—whether it represents individuals' issue involvement, the strength of controversy, or other factors—under such critical conditions, accurately inferring actual public sentiments from expressed opinions becomes increasingly challenging, as discussed by Manfredi et al. (2020).

The current simulation results further elucidate that the spiral of silence can be viewed as a specific form of opinion convergence, where expressed opinions gravitate toward majority views, while a substantial portion of population opt to conceal their true beliefs. This underscores that the spiral of silence is not an isolated phenomenon, but rather a specific pattern within a broader spectrum of intermediate states—where public opinion appears increasingly convergent, yet discrepancies between public expressions and private attitudes continue to exist. While this simulation revealed no instances where attitude polarization exceeded opinion polarization—making it unlikely that public opinion converges while highly polarized attitudes remain hidden—this does not imply that unexpressed attitudes are uniformly moderate. On the contrary, the results indicate that silent attitudes may be more dispersed than commonly assumed. This suggests that the spiral of silence may arise not only from the suppression of moderate or ambivalent views, but also from the concealment of strong or extreme attitudes, a phenomenon referred to as 'loud silence', in which expressive asymmetries obscure the true extent of attitude diversity (see Figure 2b and Appendix Figure A3).

The simulation results also show that as individuals' social reach expands—thereby increasing their average network degree—the polarization of expressed opinions tend to decrease, which in turn reduces the divergence between attitudes and opinions. This effect likely arises because larger networks expose individuals to a wider array of opinions beyond their immediate neighborhoods, reducing the average path length to distant others (see Table 1) and creating more diverse and less homogeneous opinion climates. This diversification, in turn, elevates expression thresholds, potentially limiting the visibility of extreme opinions and fostering further convergence as individuals adjust their attitudes to align with the prevailing sentiments. This finding is particularly notable given that the simulation incorporates the assumption that individuals may inherently resist opposing viewpoints, as modeled by the attraction-repulsion rule (Axelrod et al., 2021). There is an ongoing debate regarding the impact of exposure to divergent views on social polarization—Mark Zuckerberg, CEO of Meta, has claimed that presenting people with conflicting opinions could exacerbate polarization, whereas Jack Dorsey, former CEO of twitter, has posited the opposite (Keijzer et al., 2024).

Our simulation results contribute to this discussion, suggesting that even when individuals have the option to distance themselves from opposing views, increased exposure to a diversity of perspectives can indeed help in reducing polarization. This effect may stem from larger and more clustered networks, where the increased likelihood of repeated exposure to others' opinions facilitate the accumulation of social influence (Centola, 2018; Centola & Macy, 2007), which helps counterbalance the repelling forces of opposing viewpoints. It is important to note, however, that the current discussion does not consider the possibility that individuals may avoid exposure to opposing views through selective exposure or network reorganization, which could complicate the long-term dynamics between social networks and polarization.

While increased social reach is shown to reduce polarization, the current simulations reveal that this depolarizing effect is contingent on the nature of elites—a finding largely consistent with prior empirical research (Bäck et al., 2023; Banda & Cluverius, 2018). Moderate elites promote attitudinal convergence, which, when combined with network enlargement, further mitigates the polarization of attitudes and opinions. In contrast, extreme elites intensify attitude spread and polarization, counteracting the depolarizing effects of increased social reach (see Appendix Figure A4). These findings highlight the interplay of two distinct mechanisms: the enlargement of individual networks generally reduces polarization, whereas extreme elites hinder—and moderate elites enhance—the depolarizing effects of network enlargement.

In the simulations conducted for this study, no specific attributes related to the capabilities or skills of elites were modeled beyond the fact that they were individual agents endowed with larger social networks and attitudes that are always publicly disclosed. Once introduced into the network, these elites become locatable reference points that other, non-elite individuals can identify and align their attitudes with, or against. In this sense, the elites in these simulations function less as active shapers of public opinion and more as social focal points or 'magnets' that elicit reactions from non-elites. Their impact on public sentiment stems not so much from their active engagement, as traditionally understood, but from how their mere presence sets the tone for non-elites to either rally around or distance themselves from. The dynamics of public opinion, therefore, are significantly shaped by the collective response of non-elites to these social magnets.

This sheds light on a crucial, yet often overlooked, aspect of the role that influential individuals play in shaping public opinion dynamics. For many years, the role of elites in shaping public opinion has been a focal point of academic inquiry, but the emphasis has largely been on the capabilities and characteristics of these elites. Traditional scholarship often highlights the centrality of these elites in social networks—measured, for instance, by their follower counts—as well as their aptitude for persuasive communication. Yet, this approach can inadvertently downplay the agency and role of ordinary individuals, or non-elites, in the dynamics of opinion formation. Watts and Dodds (2007) punctuated this oversight in their seminal work, arguing that 'large cascades of influence are driven not by influentials but by a critical mass of easily influenced individuals' (p. 441). This suggests that it is not merely the capability of elites that alters the landscape of public opinion, but rather how the non-elites perceive and interact with these elites. It is the mere presence of elites that can serve as a catalyst for change in public opinion, as the broader population can either be attracted to or repelled by them.

Overall, these findings offer a potential explanation for the polarization frequently observed in social media discussions. By bridging weak social ties, social media substantially expands individuals' personal networks and exposes them to more diverse opinion environments. At the same time, it reduces the average path length between individuals, increasing access to influential figures (Budrikis, 2023). As people assimilate or distance their attitudes in response to extreme elites, their decisions to speak out or remain silent can inadvertently contribute to more polarized opinion climates—creating the illusion of a deeply divided public. This distorted perception may lead individuals to overestimate the extent of societal disagreement, reinforcing the belief that reconciliation is impossible and increasing hostility toward others whose views may not be so different. Without intervention, this dynamic can escalate, as moderates withdraw or shift toward more extreme positions, eventually turning the illusion into a self-fulfilling reality. To counteract this, platforms and institutions could implement algorithmic strategies that prioritize exposure to diverse—often opposing—perspectives, particularly from moderate and credible sources, helping to offset the disproportionate influence of extreme voices.

Limitations

Computer simulations inevitably have limitations when it comes to capturing the full complexity of real-world societies. Beyond the size of personal networks considered in this study, first, other network-structural elements—such as network positions, average distance, clustering, and community structures—may also play distinct roles. For instance, recent empirical evidence suggests that network positions significantly influence individuals' social behaviors, including protest participation (Larson et al., 2019). The cohesiveness of local networks could also shape how individuals respond to changes in opinion climates. Experiencing minority opinion status may vary significantly depending on whether one belongs to a densely connected, close-knit network or a more loosely organized community (Cabrera et al., 2021). Additionally, the presence of long-range connections, which significantly reduces average path lengths between individuals, could have a critical influence on the dynamics of attitudes and opinions, which may be more effectively illuminated through the application of small-world networks (Flache & Macy, 2011).

Further, this study did not account for the possibility of individuals engaging in selective exposure or reorganizing their social ties in response to shifts in the opinion climate. While such reorganization of ties may occur less frequently than short-term attitude adjustments (Lazer et al., 2010), strategies like selective media exposure and algorithmically curated content may exert a cumulative influence on the dynamics of opinions and attitudes (Prior, 2013). Moreover, individuals actively shape their networks over time, potentially creating feedback loops between perceived opinion climates and tie selection. Incorporating such mechanisms could introduce another layer of complexity to the dynamics explored here, particularly in relation to echo chamber effects and the co-evolution of network structure and opinion dynamics.

Second, the model simplifies the role of elites or opinion leaders by characterizing them solely based on their larger networks and fixed attitudes. In reality, elites may possess a range of skills and attributes, including their ability to persuade others. While the current simulations cast elites as social magnets that attract or repel attitudes, they could be more proactive agents in shaping public opinion. Furthermore, elites often operate within interconnected network among themselves, facilitating information exchange, mutual influence and reinforcement—a factor not accounted for in the present simulations. Also, the role of elites is simplified to the size of their networks in this study, neglecting other positional advantages they often hold in actual social networks. Elites may not only have extensive direct connections but could also be positioned to serve as bridges or brokers between various network clusters, including other elites (Burt, 2004). Such positional advantages can enhance their reach and influence, potentially introducing different dynamics in the formation and expression of attitudes and opinions that were not captured in the current simulations.

Third, attitude-opinion decoupling may arise not only from contextual dynamics but also from individuals' strategic decisions to maintain a degree of ambiguity or ambivalence. Particularly in controversial contexts, individuals might opt for covert signaling of their positions, rather than remaining silent or expressing them overtly to minimize associated risks (Smaldino & Turner, 2022). If individuals with strong attitudes adopt such covert signaling strategies, they could influence others's attitudes without overtly expressing their opinions, thereby introducing different, and more complex dynamics to the process, potentially amplified by other psychological biases (Steiglechner et al., 2024). Lastly, while individuals in the current simulations were assumed to have uniform levels of tolerance, tolerance levels in reality may vary widely. In future studies, a specific distribution of tolerance levels across individuals could be assumed to add another layer of complexity.

ORCID iD

Dongyoung Sohn https://orcid.org/0000-0001-5599-0054

Funding

The author disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (*MSIT; No. 2018R1A5A7059549). *Ministry of Science and ICT.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Notes

1. In the current simulations, l is set to 1.0 and τ to 5.0, ensuring the expression threshold remains bounded within [0, 1].
2. In Bramson et al. (2016), coverage is treated as a distinct measure of polarization, where higher coverage is associated with greater polarization. However, coverage alone fails to fully capture the degree of polarization in scenarios such as uniformly distributed attitudes or clustering at bipolar extremes. For instance, a uniform distribution exhibits higher coverage than bipolar clustering, despite the latter being more representative of a polarized

state due to its concentration of attitudes at opposing extremes. This highlights the limitations of coverage as a standalone measure of polarization.

3. In 0.1% to 0.2% of simulation runs, zero variance or full coverage (i.e., coverage = 1.0) were observed, both of which drive the API to zero. Because these cases were so rare, they were retained as corner cases, but not treated as meaningful outcomes.
4. Following the methodology proposed by Hamill and Gilbert (2009, 2010) for constructing spatially grounded social networks in agent-based simulations, this artificial society adopts a population density of just under 1% (approximately 1,000 out of 103,041 cells), as suggested in their framework.
5. It is widely recognized that many large-scale networks, such as the Internet, scholarly collaborations, or neural networks, exhibit scale-free or long-tail degree distributions (Newman et al., 2001). While extended social acquaintances may sometimes follow that pattern, the smaller, denser circles of family, friends, and neighbors—with whom people regularly interact in day-to-day life—are less likely to do so. Empirical studies suggest that these 'regular' networks usually range from 10 to around 60 members, averaging about 20 (Rolfe, 2014). Given that everyday political communication most often occurs in these smaller, tighter networks, this study adopts a network generator designed to reflect shorter average path lengths, higher clustering, and greater assortativity commonly observed in local social settings (Hamill & Gilbert, 2009, 2010). This approach aligns more closely with the context of frequent, local interactions emphasized in this study.
6. The simulation was repeated 20 times for each parameter combination due to the computational intensity of each run. While a larger number of replications could improve statistical robustness, 20 replications were deemed sufficient to capture general trends and variability within the computational constraints. The simulation was terminated after 1,000 updates because the model does not always converge to a single equilibrium. Instead, it often exhibits dynamic patterns or fluctuations depending on the parameter settings. Additional tests with extended runs on a subset of simulations showed no major changes in the results beyond 1,000 steps.

References

Assenova, V. A. (2018). Modeling the diffusion of complex innovations as a process of opinion formation through social networks. *PLoS One*, *13*(5), e0196699. https://doi.org/10.1371/journal.pone.0196699

Axelrod, R., Daymude, J. J., & Forrest, S. (2021). Preventing extreme polarization of political attitudes. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(50), e2102139118. https://doi.org/10.1073/pnas.2102139118

Bäck, H., Carroll, R., Renström, E., & Ryan, A. (2023). Elite communication and affective polarization among voters. *Electoral Studies*, *84*, 102639. https://doi.org/10.1016/j.electstud.2023.102639

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115

Baldassarri, D., & Bearman, P. (2007). Dynamics of political polarization. *American Sociological Review*, *72*(5), 784–811. https://doi.org/10.1177/000312240707200507

Banda, K. K., & Cluverius, J. (2018). Elite polarization, party extremity, and affective polarization. *Electoral Studies*, *56*, 90–101. https://doi.org/10.1016/j.electstud.2018.09.009

Banisch, S., & Olbrich, E. (2019). Opinion polarization by learning from social feedback. *Journal of Mathematical Sociology*, *43*(2), 76–103. https://doi.org/10.1080/0022250X.2018.1517761

Bramson, A., Grim, P., Singer, D. J., Fisher, S., Berger, W., Sack, G., & Flocken, C. (2016). Disambiguation of social polarization concepts and measures. *Journal of Mathematical Sociology*, *40*(2), 80–111. https://doi.org/10.1080/0022250X.2016.1147443

Broido, A. D., & Clauset, A. (2019). Scale-free networks are rare. *Nature Communications*, *10*, 1017. https://doi.org/10.1038/s41467-019-08746-5

Budrikis, Z. (2023). 25 years of small-world network theory. *Nature Reviews Physics*, *5*, 440–440. https://doi.org/10.1038/s42254-023-00628-6

Burt, R. (2004). Structural holes and good ideas. *American Journal of Sociology*, *110*(2), 349–399. https://doi.org/10.1086/421787

Cabrera, B., Ross, B., Röchert, D., Brünker, F., & Stieglitz, S. (2021). The influence of community structure on opinion expression: An agent-based model. *Journal of Business and Economics*, *91*(9), 1331–1355. https://doi.org/10.1007/s11573-021-01064-7

Castellano, C., Fortunato, S., & Loreto, V. (2009). Statistical physics of social dynamics. *Reviews of Modern Physics*, *81*(2), 591–646. https://doi.org/10.1103/RevModPhys.81.591

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science*, *329*(5996), 1194–1197. https://doi.org/10.1126/science.1185231

Centola, D. (2018). *How Behavior spreads: The Science of complex contagions*. Princeton University Press.

Centola, D., & Macy, M. (2007). Complex contagions and the weakness of long ties. *American Journal of Sociology*, *113*(3), 702–734. https://doi.org/10.1086/521848

Chen, H.-T. (2018). Spiral of silence on social media and the moderating role of disagreement and publicness in the network: Analyzing expressive and withdrawal behaviors. *New Media & Society*, *20*(10), 3917–3936. https://doi.org/10.1177/1461444818763384

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*, 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(9), e2023301118. https://doi.org/10.1073/pnas.2023301118

Clifford, P., & Sudbury, A. (1973). A model for spatial conflict. *Biometrika*, *60*(3), 581–588. https://doi.org/10.2307/2335008

Deffuant, G., Neau, D., Amblard, F., & Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems*, *03*(01n04), 87–98. https://doi.org/10.1142/S0219525900000078

Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., & Ryan, J. B. (2021). Affective polarization, local contexts, and public opinion in America. *Nature Human Behaviour*, *5*, 28–38. https://doi.org/10.1038/s41562-020-01012-5

Flache, A., & Macy, M. W. (2011). Small worlds and cultural polarization. *Journal of Mathematical Sociology*, *35*(1–3), 146–176. https://doi.org/10.1080/0022250X.2010.532261

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, *80*(S1), 298–320. https://doi.org/10.1093/poq/nfw006

Funk, C., Tyson, A., Kennedy, B., & Pasquini, G. (2023). *What Americans think about COVID-19 vaccines*. Pew Research Center. https://www.pewresearch.org/science/2023/05/16/what-americans-think-about-covid-19-vaccines/

Galam, S. (1997). Rational group decision making: A random field Ising model at T = 0. *Physica A*, *238*(1-4), 66–80. https://doi.org/10.1016/S0378-4371(96)00456-6

Galesic, M., Olsson, H., Dalege, J., van der Does, T., & Stein, D. L. (2021). Integrating social and cognitive aspects of belief dynamics: Towards a unifying framework. *Journal of The Royal Society Interface*, *18*(176), 1–12. https://doi.org/10.1098/rsif.2020.0857

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, *78*(6), 1360–1380. https://doi.org/10.1086/225469

Hamill, L., & Gilbert, N. (2009). Social circles: A simple structure for agent-based social network models. *Journal of Artificial Societies and Social Simulation*, *12*(2) http://jasss.soc.surrey.ac.uk/12/2/3.html

Hamill, L., & Gilbert, N. (2010). Simulating large social networks in agent-based models: A social circle model. *Emergence: Complexity and Organization*, *12*, 78–94. https://doi.org/10.1002/9781118974414

Hayes, A. F. (2007). Exploring the forms of self-censorship: On the spiral of silence and the use of opinion expression avoidance strategies. *Journal of Communication*, *57*(4), 785–802. https://doi.org/10.1111/j.1460-2466.2007.00368.x

Hayes, A. F., Glynn, C. J., & Shanahan, J. (2005). Willingness to self-censor: A construct and measurement tool for public opinion research. *International Journal of Public Opinion Research*, *17*(3), 298–323. https://doi.org/10.1093/ijpor/edh073

Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, *5*(3), 2.

Katz, D., & Allport, F. H. (1931). *Student Attitudes*. Craftsman.

Keijzer, M., Mäs, M., & Flache, A. (2024). Polarization on social media: Micro-level evidence and macro-level implications. *Journal of Artificial Societies and Social Simulation*, *27*(1), 7. https://doi.org/10.18564/jasss.5298

Kendal, R. L., Boogert, N. J., Rendell, L., Laland, K. N., Webster, M., & Jones, P. L. (2018). Social learning strategies: Bridge-building between fields. *Trends in Cognitive Sciences*, *22*(7), 651–665. https://doi.org/10.1016/j.tics.2018.04.003

Larson, J. M., Nagler, J., Ronen, J., & Tucker, J. A. (2019). Social networks and protest participation: Evidence from 130 million Twitter users. *American Journal of Political Science*, *63*(3), 690–705. https://doi.org/10.1111/ajps.12436

Latané, B., & Wolf, S. (1981). The social impact of majorities and minorities. *Psychological Review*, *88*(5), 438–453. https://doi.org/10.1037/0033-295X.88.5.438

Lazer, D., Rubineau, B., Chetkovich, C., Katz, N., & Neblo, M. (2010). The coevolution of networks and political attitudes. *Political Communication*, *27*, 248–274. https://doi.org/10.1080/10584609.2010.500187

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the “post-truth” era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008

Li, L., Fan, Y., Zeng, A., & Di, Z. (2019). Binary opinion dynamics on signed networks based on Ising model. *Physica A*, *525*(1), 433–442. https://doi.org/10.1016/j.physa.2019.03.011

Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2022). A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour*, *7*, 74–101. https://doi.org/10.1038/s41562-022-01460-1

Macy, M. W., Kitts, J. A., Flache, A., & Benard, S. (2003). Polarization in dynamic networks: A Hopfield model of emergent structure. *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, 162–173. https://doi.org/10.17226/10735

Macy, M. W., & Willer, R. (2002). From factors to actors: Computational sociology and agent-based modeling. *Annual Review of Sociology*, *28*, 143–166. https://doi.org/10.1146/annurev.soc.28.110601.141117

Manfredi, R., Guazzini, A., Roos, C. A., Postmes, T., & Koudenburg, N. (2020). Private-public opinion discrepancy. *PLoS One*, *15*(11), e0242148. https://doi.org/10.1371/journal.pone.0242148

Matthes, J., Knoll, J., & von Sikorski, C. (2018). The "spiral of silence" revisited: A meta-analysis on the relationship between perceptions of opinion support and political opinion expression. *Communication Research*, *45*(1), 3–33. https://doi.org/10.1177/0093650217745429

Matthes, J., Rios Morrison, K., & Schemer, C. (2010). A spiral of silence for some: Attitude certainty and the expression of political minority opinions. *Communication Research*, *37*(6), 774–800. https://doi.org/10.1177/0093650210362685

Miller, J. H., & Page, S. E. (2007). *Complex adaptive systems: An Introduction to computational models of Social Life*. Princeton University Press.

Moscovici, S. (1976). *Social influence and Social Change*. Academic Press.

Mutz, D. C. (2002). The consequences of cross-cutting networks for political participation. *American Journal of Political Science*, *46*(4), 838–855. https://doi.org/10.2307/3088437

Myers, D. G., & Bishop, G. D. (1970). Discussion effects on racial attitudes. *Science*, *169*(3947), 778–779. https://doi.org/10.1126/science.169.3947.778

Neuwirth, K., Frederick, E., & Mayo, C. (2007). The spiral of silence and fear of isolation. *Journal of Communication*, *57*(3), 450–468. https://doi.org/10.1111/j.1460-2466.2007.00352.x

Newman, M., Strogatz, S., & Watts, D. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, *64*, 026118. https://doi.org/10.1103/PhysRevE.64.026118

Noelle-Neumann, E. (1974). The spiral of silence: A theory of public opinion. *Journal of Communication*, *24*, 43–51. https://doi.org/10.1111/j.1460-2466.1974.tb00367.x

Noelle-Neumann, E. (1993). *The Spiral of Silence: Public Opinion—Our social skin*. University of Chicago Press.

Nordbrandt, M. (2023). Affective polarization in the digital age: Testing the direction of the relationship between social media and users' feelings for out-group parties. *New Media & Society*, *25*, 3392–3411. Advanced Online Publication. https://doi.org/10.1177/14614448211044393

Nowak, A., Szamrej, J., & Latané, B. (1990). From private attitude to public opinion: A dynamic theory of social impact. *Psychological Review*, *97*(3), 362–376. https://doi.org/10.1037/0033-295X.97.3.362

Prior, M. (2013). Media and political polarization. *Annual Review of Political Science*, *16*, 101–127. https://doi.org/10.1146/annurev-polisci-100711-135242

Prislin, R. (2022). Minority influence: An agenda for study of social change. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.911654

Robison, J., & Mullinix, K. J. (2016). Elite polarization and public opinion: How polarization is communicated and its effects. *Political Communication*, *33*(2), 261–282. https://doi.org/10.1080/10584609.2015.1055526

Röchert, D., Cargnino, M., & Neubaum, G. (2022). Two sides of the same leader: An agent-based model to analyze the effect of ambivalent opinion leaders in social networks. *Journal of computational social science*, *5*, 1159–1205. https://doi.org/10.1007/s42001-022-00161-z

Rolfe, M. (2014). Social networks and agent-based modeling. In G. Manzo (Ed.), *Analytical Sociology* (pp. 233–260). John Wiley.

Ross, B., Pilz, L., Cabrera, B., Brachten, F., Neubaum, G., & Stieglitz, S. (2019). Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems*, *28*(4), 394–412. https://doi.org/10.1080/0960085X.2018.1560920

Ross, L., Greene, D., & House, P. (1977). The “false consensus effect”: An ego-centric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*, 279–301. https://doi.org/10.1016/0022-1031(77)90049-X

Sassenberg, K., & Jonas, K. J. (2007). Attitude change and social influence on the net. In A. N. Joinson, K. Y. A. McKenna, T. Postmes, & U.-D. Reips (Eds.), *The Oxford handbook of Internet psychology* (pp. 253–271). Oxford University Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 623–656. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x

Sherif, M., & Hovland, C. I. (1961). *Social judgment: Assimilation and contrast effects in communication and attitude change*. Yale University Press.

Slater, M. D. (2007). Reinforcing spirals: The mutual influence of media selectivity and media effects and their impact on individual behavior and social identity. *Communication Theory*, *17*, 281–303. https://doi.org/10.1111/j.1468-2885.2007.00296.x

Smaldino, P. E., & Turner, M. A. (2022). Covert signaling is an adaptive communication strategy in diverse populations. *Psychological Review*, *129*(4), 812–829. https://doi.org/10.1037/rev0000344

Sobol′, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, *55*(1-3), 271–280. https://doi.org/10.1016/S0378-4754(00)00270-6

Sohn, D. (2022). Spiral of silence in the social media era: A simulation approach to the interplay between social networks and mass media. *Communication Research*, *49*(1), 139–166. https://doi.org/10.1177/0093650219856510

Sohn, D., & Choi, Y. S. (2023). Silence in social media: A multilevel analysis of the network structure effects on participation disparity in Facebook. *Social Science Computer Review*, *41*(5), 1767–1790. https://doi.org/10.1177/08944393221117917

Sohn, D., & Geidner, N. (2016). Collective dynamics of the spiral of silence: The role of ego-network size. *International Journal of Public Opinion Research*, *28*(1), 25–45. https://doi.org/10.1093/ijpor/edv005

Song, H., & Boomgaarden, H. G. (2017). Dynamic spirals put to test: An agent-based model of reinforcing spirals between selective exposure, interpersonal networks, and attitude polarization. *Journal of Communication*, *67*(2), 256–281. https://doi.org/10.1111/jcom.12290

Steiglechner, P., Keijzer, M. A., E Smaldino, P., Moser, D., & Merico, A. (2024). Noise and opinion dynamics: How ambiguity promotes pro-majority consensus in the presence of confirmation bias. *Royal Society Open Science*, *11*, 231071. https://doi.org/10.1098/rsos.231071

Sunstein, C. R. (2017). *# Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.

Thiele, J. C., Kurth, W., & Grimm, V. (2014). Facilitating parameter estimation and sensitivity analysis of agent-based models: A cookbook using NetLogo and R. *Journal of Artificial Societies and Social Simulation*, *17*(3), 11.

Tormala, Z. L., & Petty, R. E. (2004). Resistance to persuasion and attitude certainty: The moderating role of elaboration. *Personality and Social Psychology Bulletin*, *30*(11), 1446–1457. https://doi.org/10.1177/0146167204264251

Törnberg, P. (2022). How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences*, *119*(42), e2207159119. https://doi.org/10.1073/pnas.2207159119

Waldherr, A. (2014). Emergence of news waves: A social simulation approach. *Journal of Communication*, *64*, 852–873. https://doi.org/10.1111/jcom.12117

Watts, D. J., & Dodds, P. S. (2007). Influentials, networks, and public opinion formation. *Journal of Consumer Research*, *34*(4), 441–458. https://doi.org/10.1086/518527

Wilensky, U. (1999). *Netlogo. Center for Connected Learning and computer-based modeling*. Northwestern University. http://ccl.northwestern.edu/netlogo/

Author Biography

Dongyoung Sohn (Ph.D., The University of Texas at Austin) is a professor in the Department of Media and Communication at Hanyang University in Seoul, South Korea. His research interests include media psychology and computational approaches to the study of communication networks and social dynamics.

Appendix

Additional Note on Attitude Change Model

With the attraction-repulsion model for attitude change, an individual's attitude is updated as follows. Consider, for example, a person with $a_i^{(t)} = 0.7$ who is exposed to a neighbor j 's expressed opinion $o_j = 0.6$. If the attitude distance $d_{ij}^{(t)} = 0.1$ is smaller than a tolerance range ϵ, $a_i^{(t+1)}$ adjusts closer to o_j, calculated as 0.7 + (0.29) (0.65–0.7)=0.6855. If the neighbor's $o_j = 0.3$ and the distance $d_{ij}^{(t)} = 0.4$ is greater than ϵ, $a_i^{(t+1)}$ shifts away, calculated as 0.7 − (0.29) (0.5–0.7) (0.3)=0.7174. Note that there is a discontinuous jump from maximum convergence to repulsion at the tolerance threshold ϵ , reflecting a sharp transition in the direction of attitude adjustment (see Appendix Figure A1).

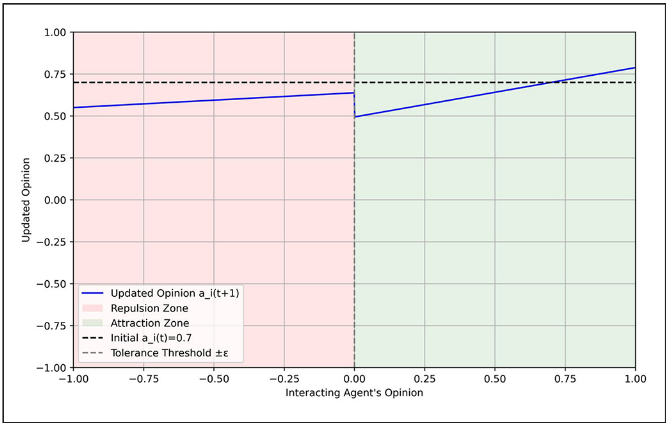


Figure A1. A visual example of attitude change.

Sensitivity Analysis

Sobol sensitivity analysis (Sobol′, 2001) has been employed to decompose the variance of model outputs into contributions attributable to input parameters and their interactions. This non-parametric sensitivity analysis technique is particularly useful for complex, nonlinear models with interdependent parameters, as it offers a comprehensive view of the influence each parameter exerts on the output variance, both individually, and in combination with others (Thiele et al., 2014). Appendix Table A1 below summarizes the analysis results and Figure A2 graphically illustrates how the model outputs were varied by the individual parameters and their interactions. First-order (S1) and total-order (ST) Sobol indices were calculated to quantify the main effects of five key parameters—tolerance, social reach, elite presence, elite diversity, elite extremity—on two output variables of the simulation—the indices for attitude polarization, and opinion polarization—and their total effects, including interactions.

For attitude polarization, the analysis revealed a relatively small but balanced contribution from most parameters, with ‘elite presence’ showing a marginally higher first-order index, suggesting a slightly more pronounced direct effect on the output. However, the closeness of the total-order indices across all parameters indicates that

Table A1. Sobol Sensitivity Analysis.

Parameters	Output variable	First-order index (S1)	First-order CI	Total-order index (ST)	Total-order CI
Tolerance	Attitude polarization	0.124	0.113, 0.135	0.021	0.018, 0.023
	Opinion polarization	0.363	0.341, 0.384	0.068	0.063, 0.073
Social reach	Attitude polarization	0.124	0.110, 0.137	0.032	0.028, 0.036
	Opinion polarization	0.579	0.542, 0.616	0.183	0.168, 0.197
Elite presence	Attitude polarization	0.129	0.114, 0.143	0.042	0.036, 0.048
	Opinion polarization	0.658	0.617, 0.700	0.236	0.218, 0.254
Elite diversity	Attitude polarization	0.003	−0.002, 0.009	0.002	−0.002, 0.007
	Opinion polarization	0.001	−0.006, 0.008	0.004	−0.003, 0.012
Elite extremity	Attitude polarization	0.127	0.114, 0.140	0.024	0.018, 0.029
	Opinion polarization	0.364	0.341, 0.387	0.072	0.063, 0.081

Note. First-order index (S1) shows the main effects of individual parameters while total-order index (ST) shows the total effects including interactions.

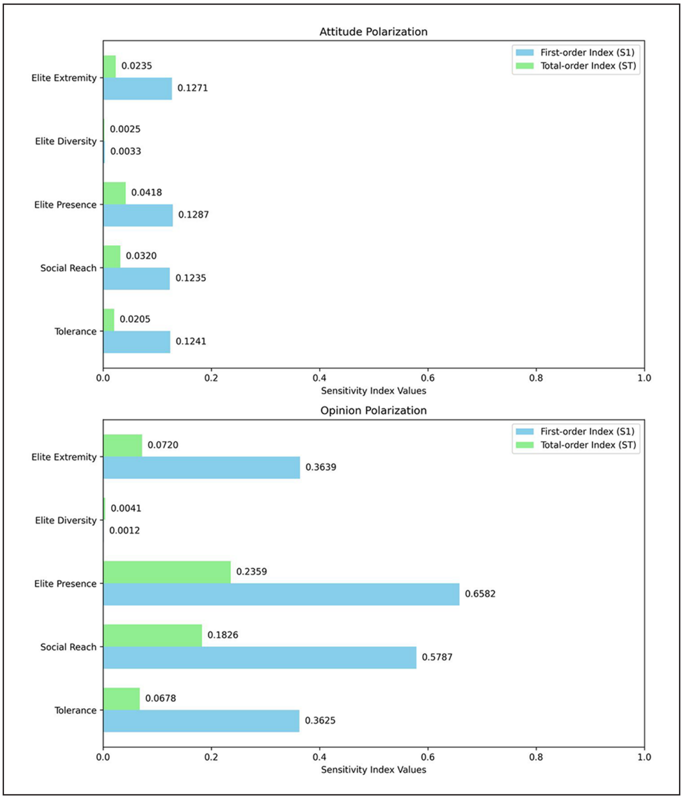


Figure A2. Bar chart for Sobol sensitivity indices.

no single parameter or set of interactions dominates the output. In contrast, opinion polarization exhibited greater variability, with ‘elite presence and ‘social reach’ both showing significantly higher first-order and total-order indices, denoting their strong individual effects and their roles in interactions with others. As depicted in Figure A2, these parameters, along with ‘tolerance’ and ‘elite extremity’, are key drivers of the variability of opinion polarization, underscoring the influence of network structure and elite characteristics on the opinion distributions within the model.

The sensitivity analysis results highlight that the distribution of attitudes remains relatively stable and less influenced by circumstantial factors, in contrast to the distribution of expressed opinions, which is more susceptible to such influences. By integratively combining attitude dynamics and opinion expression processes, the modeling approach clarifies under what conditions attitudes and expressed opinions diverge and elucidates the underlying reasons for these discrepancies. This holistic approach reveals that the public opinion landscape may appear more volatile and easily disrupted, often masking the underlying reality that people’s attitudes, thoughts, and beliefs remain relatively unchanged, consistent with the key findings from the simulations.

Additional Visualization of Individual Attitude Distributions

To further explore the dynamics of attitude change, additional visualizations of individual attitude distributions are presented. These plots highlight how attitudes evolve under varying tolerance thresholds, social connectivity, and elite influences.

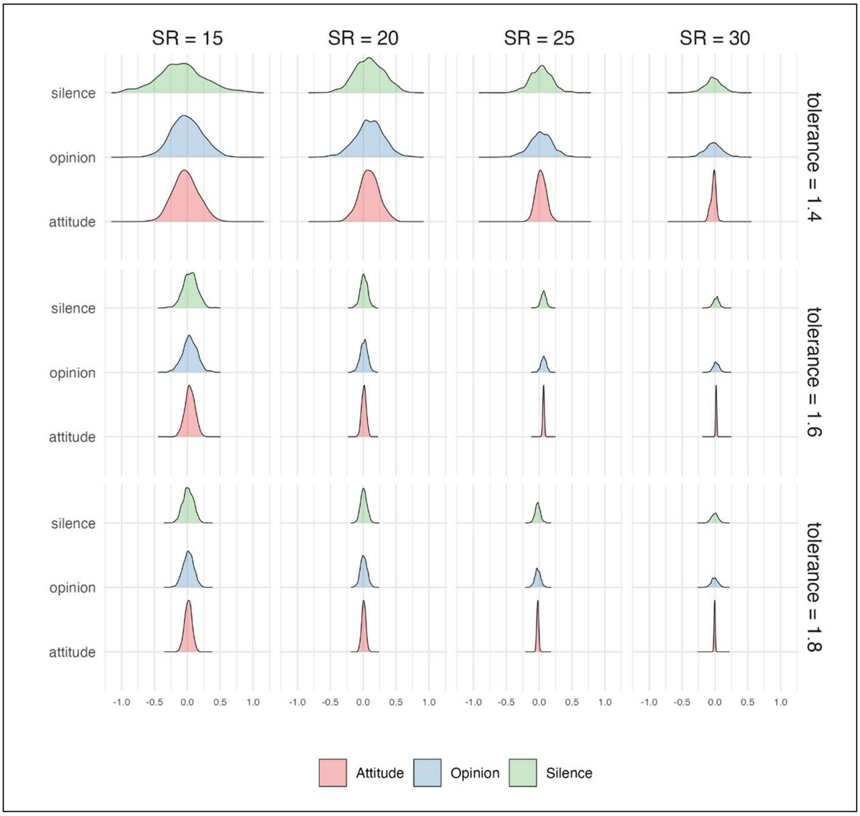


Figure A3. Attitude distributions (social reach × tolerance).

Note. These agent-level attitude distributions were captured at time = 1,000 (SR: social reach). Results are averaged over 20 repetitions for each condition.

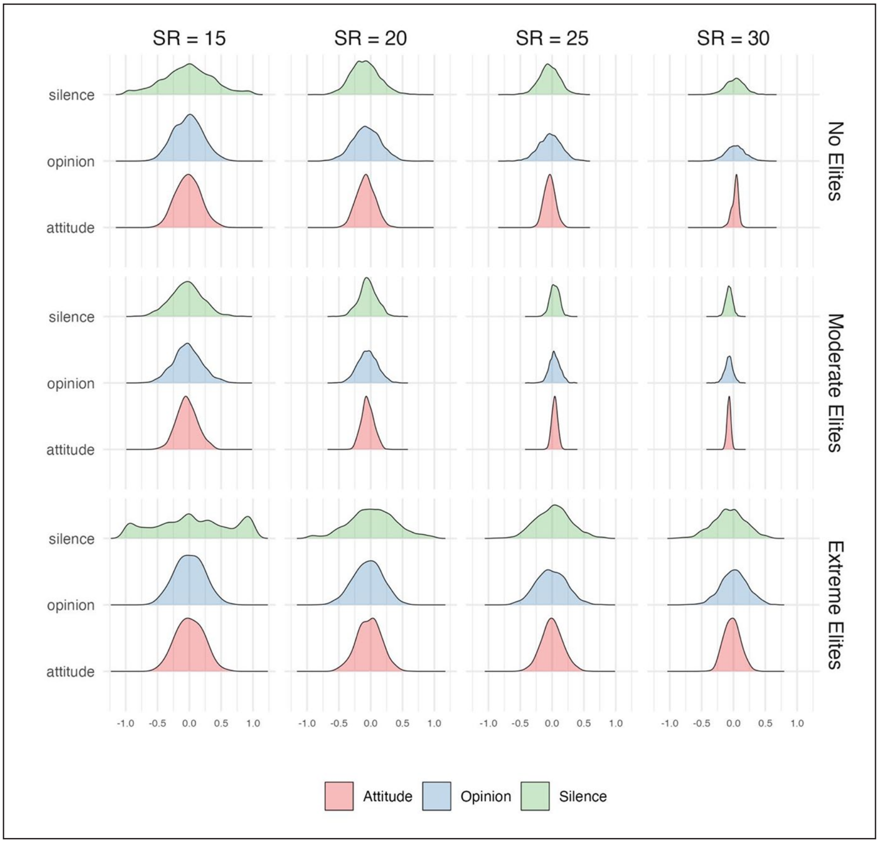


Figure A4. Attitude distributions (social reach × elite extremity).

Note. These agent-level attitude distributions were captured at time = 1,000, with the tolerance level fixed at 1.4 (SR: social reach). Results are averaged over 20 repetitions for each condition.