OPEN ACCESS
Advances in Complex Systems
Vol. 25, No. 8 (2022) 2250011 (30 pages)

DOI: 10.1142/S0219525922500114

WHEN INTUITION FAILS: THE COMPLEX EFFECTS OF ASSIMILATIVE AND REPULSIVE INFLUENCE ON OPINION POLARIZATION

SHUO LIU

School of Economics and Management,
Dalian University of Technology,
No. 2 Linggong Road, Dalian 116024, P. R. China
liushuo1260@mail.dlut.edu.cn

MICHAEL MÄS*

Institute of Technology Futures,
Karlsruhe Institute of Technology,
Douglasstraße 24, Karlsruhe 76133, Germany
michael.maes@kit.edu

HAOXIANG XIA

School of Economics and Management,
Dalian University of Technology,
No. 2 Linggong Road, Dalian 116024, P. R. China
hxxia@dlut.edu.cn

ANDREAS FLACHE

Department of Sociology, University of Groningen,
Grote Rozenstraat 31, Groningen 9712 TG, The Netherlands
a.flache@rug.nl

Received 31 October 2022
Revised 2 December 2022
Accepted 12 December 2022
Published 28 January 2023

There is a public and scholarly debate about whether personalized services of social-media platforms contribute to the rise of bipolarization of political opinions. On the one hand, it is argued that personalized services of online social networks generate filter bubbles limiting contact between users who disagree. This reduces opportunities for assimilative social influence between users from different camps and prevents opinion convergence. On the other hand, empirical research also indicated that exposing users to content from the opposite political

* Corresponding author.

spectrum can activate the counter-part of assimilative influence, repulsive influence. Fostering contact that leads to opinion assimilation and limiting contacts likely to induce repulsive interactions, it has been concluded, may therefore prevent bipolarization. With an agent-based model, we demonstrate here that these conclusions fail to capture the complexity that assimilative and repulsive influence generate in social networks. Sometimes, more assimilative influence can actually lead to more and not less opinion bipolarization. Likewise, increasing the exposure of users to like-minded individuals sometimes intensifies opinion polarization. While emerging only in specific parts of the parameter space, these counter-intuitive dynamics are robust, as our simulation experiments demonstrate. We discuss implications for the debate about filter bubbles and approaches to improve the design of online social networks. While we applaud the growing empirical research on the micro-processes of assimilative and repulsive influence in online settings, we warn that drawing conclusions about resulting macro-outcomes like opinion bipolarization requires a rigorous analysis capturing the complexity of online communication systems. Intuition alone is error-prone in this context. Accordingly, models capturing the complexity of social influence in networks should play a more important role in the design of communication systems.

Keywords: Opinion polarization; opinion dynamics; repulsion; negative influence; online social networks; filter bubbles; complexity.

1. Introduction

This paper contributes to a very recent scholarly and public debate, showing how one of the most central claims of complexity science is highly relevant for this debate: interaction on a micro-level can generate complex and counter-intuitive outcomes on a macro-scale [29, 30, 35]. The debate we address concerns the contribution of online social networks to the bipolarization of political opinions, characterized by increasing opinion differences between emergent subgroups in a population with subgroups growing internally homogeneous and mutually distinct. There is concern that personalized services of online social networks generate so-called "filter bubbles" creating information diets for users that limit exposure to content they disagree with. The lack of content challenging users' views and increased exposure to like-minded content, it is argued, intensifies users' opinions and contributes to opinion bipolarization. It has been warned that online social networks may, thus, have contributed to disruptive political events such as Brexit, the Yellow Vest movement, the 2021 Capitol riots or fierce resistance to government measures in the recent pandemic [4, 10, 23, 31–33, 36]. We demonstrate here that important contributions in this debate fail to consider the complexity arising from interaction in online social networks and, as a consequence, may draw problematic conclusions about the causes of opinion bipolarization and approaches to mitigate it.

Empirical research in the field of computational social science has made great progress in understanding the micro-processes driving opinion bipolarization on online social networks. With novel sources of data and sophisticated research designs, researchers have shown that political opinions are indeed affected by the consumption of online content [1, 8, 24]. Users align their political views to the content they consume, a process that we denote here as "assimilative influence", following [14]. There is also evidence, however, for the counter-part of assimilative influence, "repulsive

influence". In a prominent study, users describing themselves as conservative were exposed to content from liberal sources and were found to develop more conservative opinions [1]. This work resonates further empirical tests of the assumption of repulsion both in online and offline contexts. While outcomes are mixed [37], there is growing evidence for repulsive social influence in both contexts [25, 27].

These micro-level observations have led researchers to important conclusions about the contribution of online social networks to opinion bipolarization and approaches to mitigate the dynamic. It has been concluded that filter bubbles insulating users from content activating repulsion may actually help reduce opinion bipolarization. Accordingly, "simply tweaking algorithms to show partisans more content from the opposition may aggravate sectarianism rather than reducing it" [10]. Likewise, Mark Zuckerberg, founder and CEO of Facebook, argued that "ideas, like showing people an article from the opposite perspective, actually deepen polarization by framing other perspectives as foreign" [40]. In a nutshell, the underlying intuition is that assimilative influence contributes to opinion convergence and that repulsive influence increases opinion differences between users. As a consequence, algorithms preventing interaction between users that may activate repulsion and fostering instead interactions that induce assimilative influence should reduce opinion bipolarization. We show here that this intuition is incomplete.

We demonstrate that the relationship between the micro-processes of assimilative and repulsive influence, on the one hand, and opinion bipolarization, on the other hand, is more complex than intuition suggests. In particular, we show with an agent-based model that increasing individuals' openness to assimilative social influence does not necessarily lead to reduced opinion variation on the macro-level: sometimes more assimilative influence on the micro-level results in more, instead of less, opinion bipolarization. Second, we show that creating more contact between users who like each other and, thus, decreasing the relative amount of contact leading to repulsive influence can also foster rather than decrease opinion bipolarization. With regard to both results, we suggest that a "naive" intuitive reasoning might have directly projected effects of changes at the micro-level of the model to corresponding effects on the macro-level outcome of opinion bipolarization, which we show to be misleading.

To be sure, we do not argue that the macro-conclusions drawn earlier about online social networks are necessarily false. It may be true that the personalization algorithms installed on online social networks actually reduce opinion bipolarization and that bursting filter bubbles may foster rather than reduce it. However, we show that sometimes assimilative and repulsive influence have the exact opposite implication. What is more, we demonstrate these counter-intuitive implications in simple, and highly stylized settings, in order to show that even in these simple cases intuition can be flawed. This highlights that conclusions about much more complicated cases, such as online social networks, should not be based on intuition alone.

In this way, we extend a growing body of literature in complexity research developing formal models of collective opinion dynamics to explore how increasing extremism or bipolarization in a political debate can result from the complex

interplay of the interactions of multiple individuals embedded in heterogeneous social structures, without those individuals necessarily intending nor expecting their interactions to result in bipolarization or fragmentation of society [6, 11, 14, 18, 19, 26, 34]. This literature suggests that there are no easy answers to the question which impact features of online communication and structural characteristics of social networks have on opinion dynamics [23]. Yet, to our knowledge the notion that bipolarization could be tempered by more possibilities for assimilative influence between dissimilar individuals has never been challenged in this literature.

To this end, we developed an agent-based model allowing us to flexibly tweak assimilative influence and repulsion in simulated populations. Unlike earlier models where a single parameter controls the amount of assimilative influence and repulsion [5, 16, 28], we adopt here an approach proposed by Jager and Amblard [21] to control the two forces independently from each other. In addition, we included network heterogeneity, in that nodes can be connected by positive (friends) and negative (foes) social relationships. While assimilative influence can only occur between positively connected nodes in our approach, repulsive influence is only possible between negatively connected nodes. This manipulation of the characteristics of the network ties, thus, provides us with a means to manipulate structural possibilities for assimilative and repulsive influence. For instance, a network consisting of two groups with group members connected by positive ties and many negative links between the groups can be seen as a structure with a high potential for bipolarization between those groups. Yet, as we will show, increasing structural possibilities for assimilative influence relative to those for repulsive influence does not necessarily reduce bipolarization between groups. To be sure, the valence of links in our model is taken to be a static feature of the network. While one could argue that agreement or disagreement between any two agents should also affect the valence of their link, we believe that it is important to leave this complication to future work. One important reason is that we want the signed network to represent the structural side of so-called "affective" polarization or "sectarianism", a concept that has received increasing attention in recent work on bipolarization [10, 20]. Affective polarization describes a state where individuals identify strongly with a political camp and have negative affect toward other camps. Intuitively, one would expect that growing negative affect between members of different political camps fosters repulsive influence [9, 14, 17] and, as a consequence, creates a breeding ground for opinion bipolarization. As we will show in our study, this intuition turns out to be incomplete. While we are aware that affect towards other groups as such is also subject to changes in intergroup attitudes [11], we are here interested in the intergroup relations as a structural and stable phenomenon that may emerge out of a long history of positive and negative interactions within and between groups, not to be changed easily in the short term. Relations in the signed network then represent how easily assimilative or repulsive influence is triggered, where we assume that sufficient agreement to positively connected others can trigger assimilation, while sufficient disagreement to negatively connected neighbors can be the driver of further repulsion.

Section 2 summarizes our model incorporating assimilative and repulsive influence on individuals in heterogeneous networks. Results are presented in two steps. In Sec. 3.1, we analyze a highly stylized example to demonstrate that sometimes increasing individuals' openness to assimilative influence can intensify opinion bipolarization. We show with the same example that more structural possibilities for assimilative influence in a population sometimes leads to more opinion bipolarization. Section 3.2 is concerned with the robustness of these findings. We present simulation experiments testing whether the two counter-intuitive findings can be replicated when different network structures and different initial opinion distributions are assumed than in the stylized example. We discuss implications for empirical research on online social-influence processes and the debate about the effects of online social media on opinion bipolarization in Sec. 4.

2. The Model

To demonstrate that assimilative and repulsive influence can aggregate to complex and counter-intuitive macro-outcomes, we developed an agent-based model of opinion dynamics on a heterogeneous network. Agents are represented as nodes in an undirected network with N individuals. Each i agent is described by an opinion $x_i(t)$ representing i's stance on an issue at time t. Opinions are measured on a continuous scale ranging from zero to one ($0 \leq x_i(t) \leq 1$).

The network structure is fixed. Network edges are represented in a matrix C defining for all pairs ij of agents whether they have no social relationship ($c_{ij} = 0$), a positive relationship ($c_{ij} = 1$), or a negative relationship ($c_{ij} = -1$). Positive and negative relationships represent that the two agents are communicating in that they send and receive online content reflecting their opinions. That is, we do not model the emission and diffusion of content in a network [22] but assume that agents connected by a positive or a negative link are aware of each others' opinion as a result of the communication. Matrix C can be split into C^+ and C^-, which represent the positive and negative edge structure, respectively, and satisfy $C = C^+ + C^-$. Figure 1 provides an example of a network with six agents for illustration. The algorithms to generate the actual networks are described in detail below.

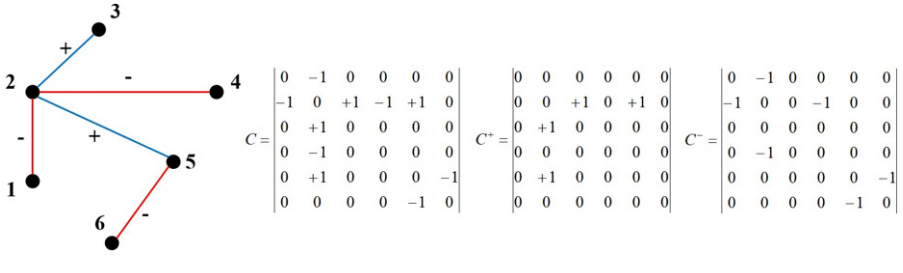
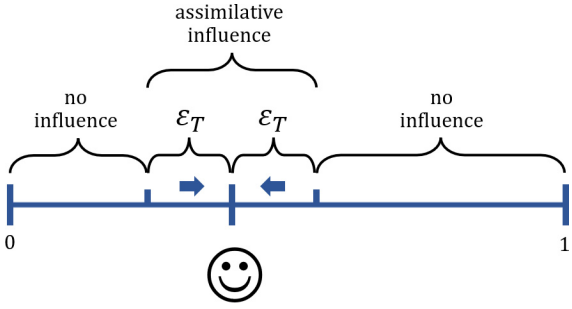
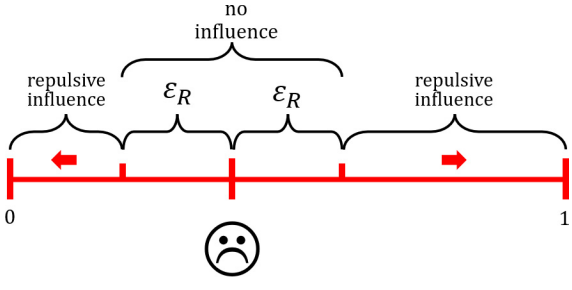


Fig. 1. Example network where nodes represent agents; the symbol +/− next to the edge indicates whether the link is positive or negative. C is the matrix of the whole network. C^+ and C^- represent the structural matrices of positive and negative links, respectively.



(a) Positive links



(b) Negative links

Fig. 2. Assimilative influence and repulsive influence as implemented in Eq. (2).

The network relationships determine how connected agents can exert influence on each other's opinions. When there is no relationship between agents i and j, then they do not influence each other. Figure 2 visualizes our core assumptions about social influence, as adopted from Jager and Amblard [21]. Panel (a) shows the opinion scale and the opinion adjustments of a target i of influence (see the smiley face) caused by the source j of influence if the two are connected by a *positive* tie. If the opinion of the source does not differ by more than a given threshold ϵ_T from the opinion of the target of influence, then the target is positively influenced. Otherwise, the target's opinion remains unaffected. ϵ_T denotes the "trust threshold": the parameter specifying the maximal opinion distance between friends that still activates assimilative influence. That is, when the opinion distance of two agents connected by a positive tie exceeds the value of the trust threshold, then there is no influence. Panel (b) shows the case of a negative tie between the two agents. Here, the target does not change its opinion if the opinions of the two agents differ by less than the threshold ϵ_R. When the opinion distance, in contrast, exceeds this threshold, the target of influence changes her opinion away from the source. ϵ_R denotes the "repulsion threshold", the minimal opinion distance between agents connected by a negative link that activates repulsion. If opinions differ by less than this threshold, then there is no influence. Note that the two parameters ϵ_T and ϵ_R are independent of

each other, which allows us to tweak the amount of assimilative influence and the amount of repulsion in the population independently.

At every time step t, all agents synchronously update their opinions, based on assimilative influence by sufficiently similar network neighbors with positive relationships and repulsion by sufficiently dissimilar neighbors with negative relationships. We use a synchronous updating of opinions, in order to exclude the effect of the ordering of opinion updates on the opinion dynamics. In fact, synchronous updating turns some of the model dynamics we study deterministic, which makes it easier to understand the counter-intuitive effects the model generates.

Formally, Eqs. (1) define two subsets of agents connected to the focal agent i. Subset $A_i^+(t)$ consists of agent i's positive relationships where assimilative influence is activated. Subset $A_i^-(t)$ contains all sufficiently dissimilar foes of i, where c_{ij}^+ indicates a positive link between individual i and individual j ($c_{ij}^+ \in C^+$) and c_{ij}^- indicates a negative link between individual i and individual j ($c_{ij}^- \in C^-$). The symbols $|A_i^+(t)|$, $|A_i^-(t)|$ denote the number of individuals in subsets $A_i^+(t)$ and $A_i^-(t)$, respectively.

$$\begin{aligned} A_i^+(t) &= \{\text{sign}(\epsilon_T - |x_i(t) - x_j(t)|) \cdot c_{ij}^+ \geq 0\}, \\ A_i^-(t) &= \{\text{sign}(|x_i(t) - x_j(t)| - \epsilon_R) \cdot c_{ij}^- \leq 0\}. \end{aligned} \tag{1}$$

Equation (2) defines the opinion update. The first term in round brackets implements assimilative influence. The second term in round brackets represents the repulsion forces acting on agent i's opinion. The term $X_i^+(t)$ in the first brackets is the average opinion of all agents j that have a positive relation with i and that are sufficiently similar. $X_i^-(t)$, accordingly, is the average opinion of all agents j connected to i with a negative relationship and a sufficiently dissimilar opinion.

$$x_i(t+1) = x_i(t) + \lambda \cdot (\mu_{i,t}^+ \cdot (X_i^+(t) - x_i(t)) + \mu_{i,t}^- \cdot (x_i(t) - X_i^-(t))). \tag{2}$$

The parameter λ controls the overall strength of the influence and is set to $\lambda = 0.5$. Parameters $\mu_{i,t}^+$ and $\mu_{i,t}^-$ control agent i's openness to social influence by similar friends and dissimilar foes. In particular, Eqs. (3) implement that the size of the opinion shifts resulting from assimilative and repulsive influence is proportional to the number of activated positive and negative links. In a nutshell, the two equations implement that positive influence on an agent i is stronger when agent i's network neighborhood consists of many positive ties and when a large share of these positive ties connect i to sufficiently similar agents. Likewise, repulsion is stronger when agent i has many negative ties and the share of negative ties connecting i to sufficiently dissimilar agents is high. N_i denotes the number of neighbors of individual i.

$$\begin{aligned} \mu_{i,t}^+ &= \frac{|A_i^+(t)|}{N_i}, \\ \mu_{i,t}^- &= \frac{|A_i^-(t)|}{N_i}. \end{aligned} \tag{3}$$

3. Results

We employ our model to demonstrate that the conjunction of assimilative and repulsive influences can generate counter-intuitive outcomes at the collective level. We first show that widening the range of conditions under which assimilative influence can occur by increasing the parameter ϵ_T sometimes results in more and not less bipolarization. Second, we focus on the effects of the share of positive network ties in the network, representing structural opportunities for assimilative influence. We show that increasing the share of positive ties can result in growing opinion bipolarization.

We proceed in two steps. First, we demonstrate the two counter-intuitive effects with a highly stylized example, in order to make clear why the model generates them (see Sec. 3.1). To this end, we assume very small populations with very simple network structures. Furthermore, the dynamics generated in these stylized examples are deterministic, which implies that they do not result from an idiosyncratic effect of randomness. Second, we test in Sec. 3.2 whether the two counter-intuitive effects demonstrate here are robust to assuming different initial opinion distributions and increasing the complexity of the underlying social network.

To quantify the central outcome variable, opinion bipolarization, we use the bipolarization index [12], slightly adapted from [11]. This measure captures the degree to which opinions in the population fall apart into two evenly sized sub-populations with maximal disagreement between and maximal agreement within the clusters. Technically, bipolarization $b_P(t)$ is the variance of the opinion distances of all pairs of agents in the population at time t. Equation (4) summarizes the calculation, where $\overline{d(t)}$ represents the average opinion difference between all agents in the entire group.

$$b_P(t) = \frac{4}{N^2} \sum_{i,j}^{i\in N, j\in N} (|x_i(t) - x_j(t)| - \overline{d(t)})^2. \tag{4}$$

When all agents hold the same opinion, the bipolarization index adopts a value of zero, since the opinion distance of all pairs of agents is exactly zero, which translates into a variance of distances of zero ($b_P(t) = 0$). In contrast, when the population consists of two equally large and maximally dissimilar subgroups, then half of the pairs of agents are characterized by a distance of zero and the other half has a distance of one. The variance of distances, accordingly, is maximal and generates a bipolarization of $b_P(t) = 1$. Bipolarization values between zero and one indicate that the population has not reached consensus but bipolarization is not maximal; either because the population consists of more than two subgroups, or because subgroups are not homogeneous, or because subgroups have not adopted extreme opinions.

3.1. *Stylized example*

3.1.1. *More assimilative influence generates more polarization*

Figure 3 shows a very simple network used as a starting point for our analyses. There are $N = 10$ agents integrated in a ring network where each agent has exactly two

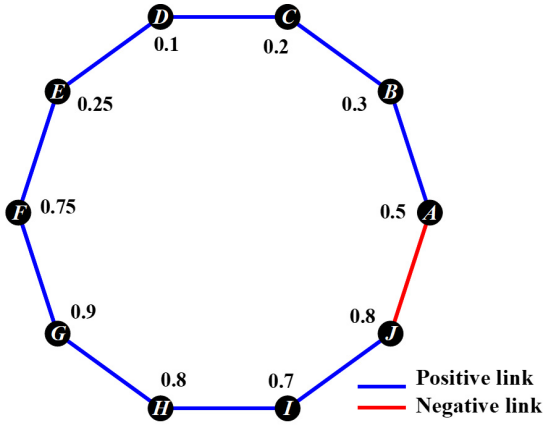


Fig. 3. (Color online) Stylized network with 10 agents. Nodes represent agents, blue links represent positive relationships (friends), and red links represent negative relations (foes). The values next to the nodes show the initial opinion assigned to the respective agent.

relationships. There are nine positive links (shown in blue). Only the link between agents A and J is negative.

The initial opinion $x_i(0)$ of each agent is shown next to the respective network node. We assigned these values in such a way that the network consists of two distinct subsets of agents to create the potential for a "group split" [13, 31]. To this end, agents A to E hold opinions equal or smaller than 0.5 and form a subset of agents with only positive ties and similar opinions. Likewise agents F to J are connected by positive ties and hold similar initial opinions. Meanwhile, their opinions adopt values above 0.5, unlike the opinions of the first subset. The two subsets are connected by one negative link between A and J and one positive link between E and F. According to Eq. (4), $\overline{d(t)} = 0.53$, the degree of bipolarization in the initial network is $b_P(0) = 0.24$.

The level of disagreement between agents connected by negative links can be critical for the emergence of opinion bipolarization in such a structure. In Fig. 3, the link A–J could induce a repulsive dynamic between these two agents if their initial disagreement is sufficiently large. If this happens, their opinions will be pushed towards opposite poles of the opinion interval. As a consequence, they can pull those connected to them via positive links (B and I, respectively) into the same direction, potentially splitting the rest of the population along the line of the initial "faultline" which separates the two subgroups. However, as we will show, whether this happens and which opinion distribution eventually arises, depends sensitively on the exact initial conditions and parameters of the model.

In Fig. 4, we show the opinion trajectories the model generates in the stylized example of Fig. 3 for three different values of the trust threshold ϵ_T. More specifically,

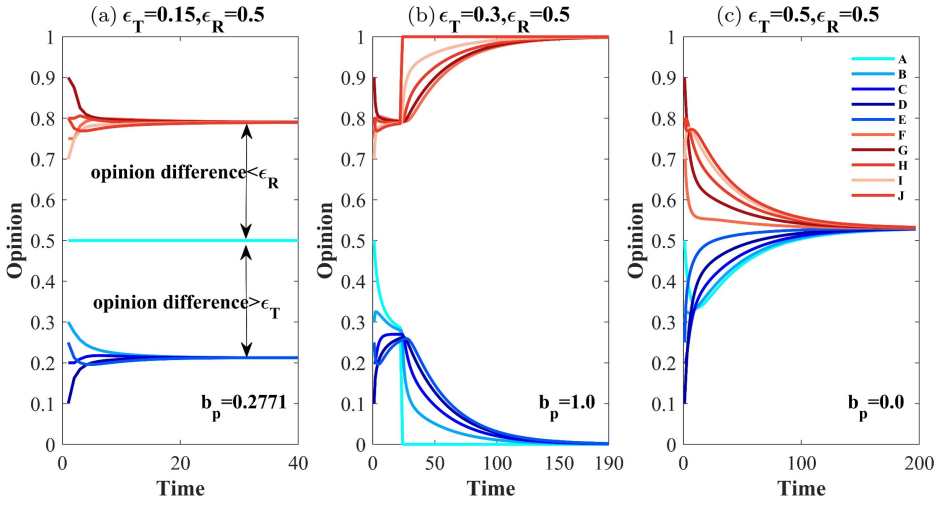


Fig. 4. Opinion trajectories generated by the model for the stylized example shown in Fig. 3 for three values of the trust threshold: (a) $\epsilon_T = 0.15$; (b) $\epsilon_T = 0.3$; and (c) $\epsilon_T = 0.5$. In all three scenarios, we fixed the repulsion threshold at $\epsilon_R = 0.5$.

for the baseline scenario we choose $\epsilon_T = 0.15$ and $\epsilon_R = 0.5$. Next, we increase ϵ_T to values of 0.3 and 0.5, respectively. Increasing the trust threshold allows more assimilative influence between agents with dissimilar opinions. Conceptually, this can be interpreted as raising the level of trust among the members of a society, which increases openness for being influenced by distant opinions of other members of the same society. Intuitively, one could expect this to support the formation of consensus and make polarized opinion distributions less likely to occur. This, however, turns out to be partly incorrect.

In the trajectory graphs, each line shows the opinion evolution of one of the 10 agents. In Panel (a) of Fig. 4, assimilative influence is activated only when nodes hold opinions differing no more than $\epsilon_T = 0.15$ from the opinions of their neighbors. The opinion differences within the upper and the bottom section of the network are small enough to lead to the formation of two local clusters with identical opinions. Importantly, the initial opinion difference between the two agents connecting the two subgroups, E and F, exceeds the trust threshold. This implies that agents E and F do not influence each other from the outset although they have a positive link. At the same time, the initial opinion differences within each of the two subgroups B–E and F–J are small enough so that assimilative influence can lead to the convergence of opinions within each of the two subgroups. Only agent A is not influenced by any of her neighbors, as A's disagreement with B is too high to cause assimilative influence and A's disagreement with J is too small to cause rejection. As a result, A retains her initial opinion. Collectively, the population thus falls apart into three clusters without repulsive influence between them.

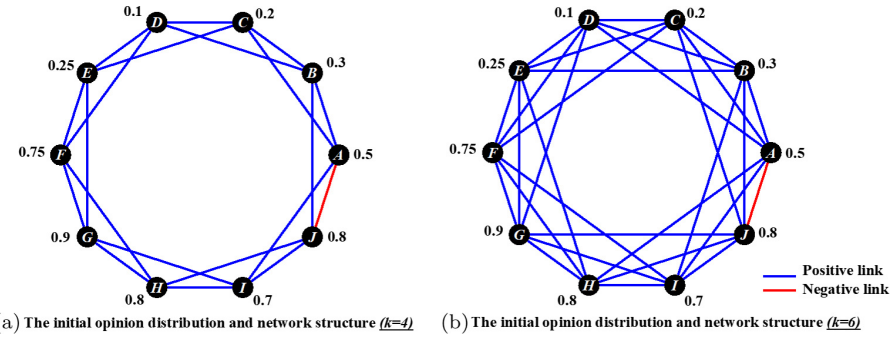
Panel (b) of Fig. 4 shows that increasing the trust threshold to $\epsilon_T = 0.3$, does not generate more but less opinion convergence than in Panel (a), contradicting the intuition nurtured above. Widening the confidence range for assimilative influence makes agent A open to influence by agent B. Accordingly, A joins the upper opinion cluster. The opinion convergence within that cluster, however, dragged A's opinion farther away from the opinion of agent J, increasing their opinion distance beyond the critical threshold of $\epsilon_R = 0.5$. Thus, repulsion is activated and A adopts increasingly extreme opinions. Next, A is exerting assimilative influence on B and, indirectly, on the remainder of the upper cluster, which drags their opinions to the lower pole of the scale. On the other side of the spectrum, agent J rejects A's opinion, pushing J to move towards the opposite pole of the opinion interval. This subsequently pulls the other agents of the bottom segment with J. In a nutshell, opinions bipolarized because increased assimilative influence generated local opinion convergence within the two clusters which, in turn, triggered repulsive influence between the clusters, eventually pushing both subgroups towards opposite extremes.

Panel (c) of Fig. 4 demonstrates that the effect of increasing the trust threshold cannot simply be extrapolated to the next level of $\epsilon_T = 0.5$. At this level, the opposite dynamic unfolds of what we observed for $\epsilon_T = 0.3$. Now, the confidence threshold is sufficiently large so that assimilative influence occurs from the outset between agents F and E. This entails the convergence of the two initial subgroups into one opinion cluster. As a consequence, also the negative relationship between A and J is not activated so that there is no repulsive influence that could prevent the emergence of a consensus.

3.1.2. *More positive links can increase opinion bipolarization*

The second counter-intuitive effect is concerned with the effects of adding positive links which provide more structural opportunities for assimilative influence both within as well as between different subgroups in the opinion space. Intuitively, this could be interpreted as bursting potential filter bubbles in an online setting before opinion dynamics could generate increasing bipolarization between otherwise disconnected subgroups, a measure that has been suggested to decrease potentials for opinion bipolarization. Again, we will show that this intuition is incomplete.

Like in the experiments reported in the previous section, we induced more possibilities for assimilative influence between agents. However, rather than increasing agents' tolerance for disagreement, we now explore what happens if more opportunities for assimilative influence are induced by changes in the share of positive links in the network. More specifically, we increased the network degree of all agents as depicted in Fig. 5, adding only positive links. The figure shows two additional experimental conditions. In addition to the network of Fig. 3 where we link each agent to the two closest neighbors on the ring network, we study networks where each agent connects with the four (Panel (a)) and the six (Panel (b)) closest neighbors. Importantly, in all three networks, there is only a single negative network



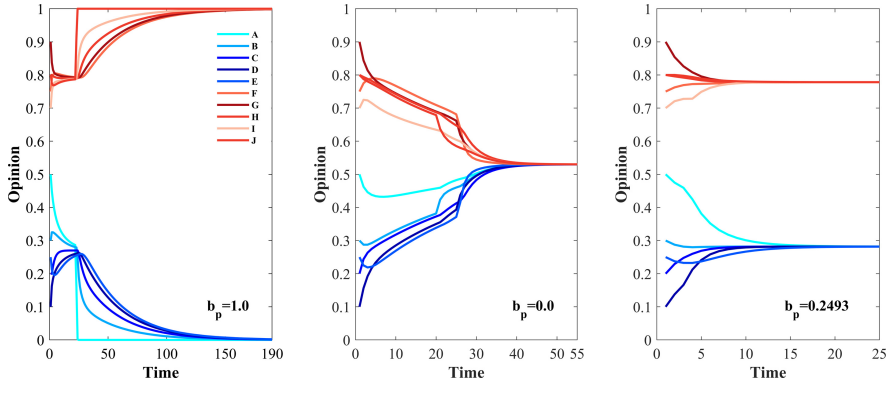
(a) **The initial opinion distribution and network structure *(k=4)*** (b) **The initial opinion distribution and network structure *(k=6)***

Fig. 5. Experimental manipulation of the share of positive links: (a) The ring network with the average degree of 4. (b) The ring network with the average degree of 6. In all networks (see also Fig. 3), there is exactly one negative link (link between A and J).

relationship connecting agents A and J. That is, increasing the network degree from 2 to 4 to 6 reduces the share of negative links in the network, greatly increasing the structural opportunities for assimilative influence relative to those for repulsive influence. While the share of negative links is 10% with $k = 2$, it drops to 5% and 2.5%, respectively, for $k = 4$ and $k = 6$.

Figure 6 shows how opinion dynamics change when the network degree k is increased from 2 to 4 to 6, respectively. We assumed the exact same initial opinions as in the stylized example reported above. Thus, also here, the network consists of two subsets of agents with similar opinions to create a potential for a group split. We kept unchanged the trust ($\epsilon_T = 0.3$) and the repulsion threshold ($\epsilon_R = 0.5$).

Note that Panel (a) in Fig. 6 is identical to Panel (b) in Fig. 4. With the average degree of two, assimilative influence leads to local convergence of opinions that pulls



(a) $k = 2$ (b) $k = 4$ (c) $k = 6$

Fig. 6. The evolution of opinions in the network when the trust threshold is 0.3 and the repulsion threshold is 0.5, and (a) the average degree of the network node is 2, (b) the average degree of the network node is 4, (c) the average degree of the network node is 6.

the two agents with the negative relationship sufficiently apart to trigger repulsive influence, resulting in sharp bipolarization at the macro-level. In line with intuition, increasing the degree to four decreases bipolarization and even generates consensus (see Panel (b) in Fig. 6). While Panel (b) still shows that opinions first converge locally within the two initial clusters, the opinions within these two clusters grow more moderate. This happens in particular because the agents with extreme initial opinions have now more relationships that influence them to adopt moderate opinions. Agent D, for instance, is now also influenced by agent B, who exerts a force towards a more moderate view compared to the influence agent D was exposed to with only two neighbors. Since extreme agents grow moderate quickly, they exert weaker pulls towards extreme views. The two clusters, therefore, adopt opinions that are not sufficiently dissimilar to activate repulsion between A and J. A, in fact, now acts as a bridge between the clusters. This never happened in the dynamics shown in Panel (a), where the assimilative force on agent A was pulling her opinion in one direction only.

Once again, increasing the possibilities for assimilative influence turns out to not have a monotonic effect on bipolarization. Panel (c) does not show the same or even stronger convergence towards consensus than Panel (b) even though even more positive relationships were added. An important reason for this effect is that only those positive relationships connecting sufficiently similar agents are effective. As a consequence, there is now even stronger local convergence within the initial clusters in the population. Thus, these initial clusters are faster in forming a local consensus which, in particular, pulls those agents faster away from the opinions of their outside neighbors, who could otherwise have served as bridges towards the other cluster. Again, agent B serves as a good example. Compared to the case in Panel (b), agent B from Panel (c) is connected to E and I, but only E is sufficiently similar to effectively influence agent B. The stronger local convergence changes the dynamics in an important way: the two clusters do now no longer grow sufficiently similar to influence each other. What is more, agent A is joining one of the clusters and, thus, no longer acts as a bridge between the two clusters. In a nutshell, in the scenario with a potential for group split that we are considering here, more assimilative influence also implies that convergence occurs more within the subgroups rather than between them.

To be sure, in the experiment presented here, effects of increasing the number of positive links cannot be cleanly disentangled from the effects this also has on the number of ties in the network. Alternatively, we could have manipulated the proportion of positive and negative links, keeping the number of links in the network constant. However, this would have changed the number of both positive and negative links, leaving again open whether changing their proportion or the numbers causes the effects observed. We believe that the choice for adding a small number of positive links while keeping the number of negative links constant is well suited to capture the counter-intuitive effect of reducing the proportion of negative links.

3.2. *Robustness analyses*

Obviously, the two counter-intuitive findings were generated in a carefully engineered setting. This allowed us to demonstrate the two effects in a setting that is not too complicated to analyze. Furthermore, the simplicity of the settings makes it possible to explain why the outcomes of the dynamics generated by assimilative and repulsive influence contradict intuition.

Next, we explore the robustness of the two counter-intuitive effects in order to test whether they occur only under the very stylized conditions assumed above or whether they can be generated also in other settings. We present our tests in two steps. First, we keep the network structure unchanged and only assume different distributions of the initial opinions. Second, we explore different network structures and larger-sized populations.

3.2.1. *Initial opinion distribution*

In order to test the robustness of our findings to changes in the initial opinion distribution, we conducted simulation experiments where we assumed the exact same network structure as shown in Figs. 3 and 5, now randomizing the initial opinions. To retain the assumption of an initial faultline potentially splitting the population, the network remained equally divided into two camps, with agents A to E in one camp having initial opinion values randomly distributed between 0 and 0.5, and agents F to J having initial opinion values randomly distributed between 0.5 and 1. All other parameters remained unchanged.

Figure 7 reports results for the first counter-intuitive effect. To generate it, we conducted a simulation experiment in which we generated 20 different initial opinion distributions and studied for each distribution the opinion dynamics for three values of the trust threshold ϵ_T. In addition, we replicated this experiment with 500 independent initial opinion distributions per treatment and report summary statistics in the text and a histogram in Appendix A.1.

Figure 7 demonstrates that our finding that increasing assimilative social influence can generate more polarization is robust to randomized initial opinions. The figure shows box plots of the bipolarization index calculated when each of the simulation runs had reached equilibrium. It shows the same qualitative pattern as Fig. 4. When the trust threshold is small, at $\epsilon_T = 0.15$, the average degree of bipolarization is above zero but low. Under this condition, dynamics typically generate multiple coexisting opinion clusters, which results in relatively low degrees of bipolarization. If agents A and J happen to hold sufficiently dissimilar initial opinions, it is possible that they adopt maximally extreme opinions due to repulsive influence, unlike in the run shown in Panel (a) of Fig. 4. However, this only happens when A and J hold relatively extreme opinions already at the outset of the dynamics. This, in turn, makes it likely that the remaining agents hold opinions that are too distant from the opinions of A and J to generate assimilative influence in this treatment of the experiment. As a consequence, the remaining agents are not attracted to A and J and high values of bipolarization are very unlikely.

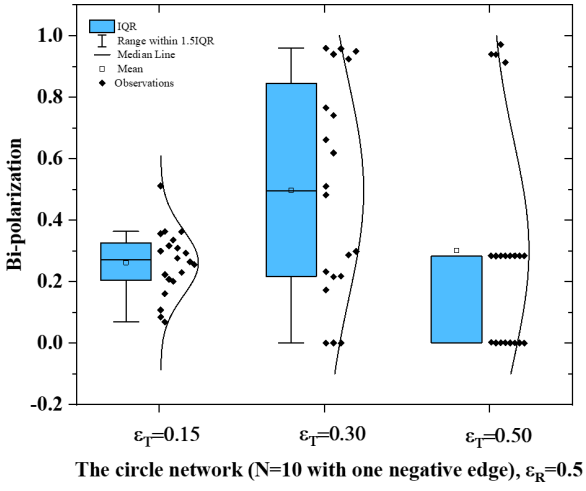


Fig. 7. (Color online) Simulation experiment testing the robustness of the counter-intuitive finding that increasing the trust threshold ϵ_T can generate more polarization. Initial opinions were randomly assigned with agents A to E having initial opinion values randomly distributed between 0 and 0.5, and agents F to J having initial opinion values randomly distributed between 0.5 and 1. In three treatments, we increased the trust threshold from $\epsilon_T = 0.15$ (a), to $\epsilon_T = 0.3$ (b), to $\epsilon_T = 0.5$ (c). Blue areas show the interquartile range (IQR). The black dots identify the observed degree of bipolarization observed in the 20 runs per treatment. For the central box, we provide in Appendix A.1 a histogram reporting bipolarization in 500 simulation runs.

Increasing the trust threshold to $\epsilon_T = 0.3$ generated a higher average degree of bipolarization, showing that the counter-intuitive finding is robust. However, we also observe more variation in bipolarization levels between runs, indicating that the initial opinion distribution has a strong influence on the opinion evolution. In Appendix A.1, we show a histogram showing results of a replication study with 500 simulation runs. A big share of the runs (45%) did end in perfect bipolarization, the equilibrium we also found in Fig. 4. Bipolarization emerges more likely when the initial opinion distance between the two agents with the negative network link (A and J) exceeds the threshold of $\epsilon_R = 0.5$, which happens in about half of the runs as a result of the opinion initialization. In addition, if the other two agents connecting the two subsets of agents (E and F) initially hold opinions differing by more than the trust threshold of $\epsilon_T = 0.3$ bipolarization is more likely. Otherwise, the assimilative influence between these two agents acts as a counter-force to opinion bipolarization. However, since the initial opinion distance between E and F very likely differs by more than 0.3, bipolarization emerges very frequently in this treatment of the experiment. When A and J happen to hold very similar opinions at the outset and when also the opinion differences among the remaining agents happen to be low, then perfect opinion consensus is possible. However, we observed consensus in fewer than 5% of the runs. In the remainder of the simulation runs under $\epsilon_T = 0.3$, we observed multiple coexisting opinion clusters similar to the pattern found in the treatment with $\epsilon_T = 0.15$. However, since there was more assimilative social influence, it was

also more likely that once A and J adopted extreme opinions due to mutual repulsive influence, other agents were attracted to these agents and also adopted extreme views. This implies that more agents joined the very extreme clusters, which translates into higher values of bipolarization than under $\epsilon_T = 0.15$.

Increasing the trust threshold further to $\epsilon_T = 0.5$, leads to lower bipolarization on average but again there is variation. First, about 35% of the runs did end in a state of perfect consensus, which is the same outcome as shown in Fig. 4. Consensus is the typical outcome when the opinions of agents A and J differ by less than $\epsilon_R = 0.5$, which happens in about half of the runs as a result of the opinion initialization. Second, about 30% of the runs end in a state of perfect bipolarization into two subgroups. Bipolarization is a likely outcome when the initial opinions of agents A and J differ by more than $\epsilon_R = 0.5$, since their negative network link generates repulsive influence under this condition. A second driver of bipolarization is the initial opinion distance between agents E and F, the two agents who establish the second network connection between the two subsets of agents. If their initial opinions differ by more than $\epsilon_T = 0.5$, then there is no opinion convergence between the two. In addition, they will pull their subgroup members towards their initial opinions and can, indirectly, pull the opinion of agents A and J so far away from each other that repulsive influence is activated and opinion polarization can emerge. The third possible equilibrium is very interesting. In about 30% of the runs, opinions are perfectly scattered across the opinion space in equilibrium with agents A and J adopting maximally extreme opinions and the remaining agents' opinions perfectly scattered with an opinion distance of $1/9$ to their closest network neighbors. This is a fixed state, because the positive network links pull agents with equal force into opposite directions, which aggregates to zero opinion change. This scattered state is the most frequent outcome when the initial opinions of A and J differ by more than $\epsilon_R = 0.5$ and if the opinions of E and F differ by less than $\epsilon_T = 0.5$. In the other two treatments, this steady state is also possible, but much less likely. Obviously, this equilibrium is an artifact of the very symmetric network structure assumed in this stylized example. However, it serves as another demonstration that even in seemingly simple cases, the conjunction of assimilative and repulsive influence generates complex outcomes.

To further assess the robustness of the nonmonotonous effects of increasing the share of positive links, we replicated the experiment shown in Fig. 5, now again with randomized initial opinions generated in the same way than reported above. As a further check, we also investigated the effect of the trust threshold and the repulsion threshold on bipolarization.

Figure 8 shows phase diagrams reporting the bipolarization value averaged across 20 independent realizations per experimental treatment. Light(Yellow) cells show that runs ended in a state of bipolarization. Dark(Blue) cells, in contrast, indicate that runs ended in a state of perfect consensus. We observe in all three panels a bipolarization phase, showing that even in the treatment where only 3.3% of the links were negative ($k = 6$), opinion distributions bipolarized. In line with the intuitive prediction that a smaller share of negative ties in the network, the diagrams reveal

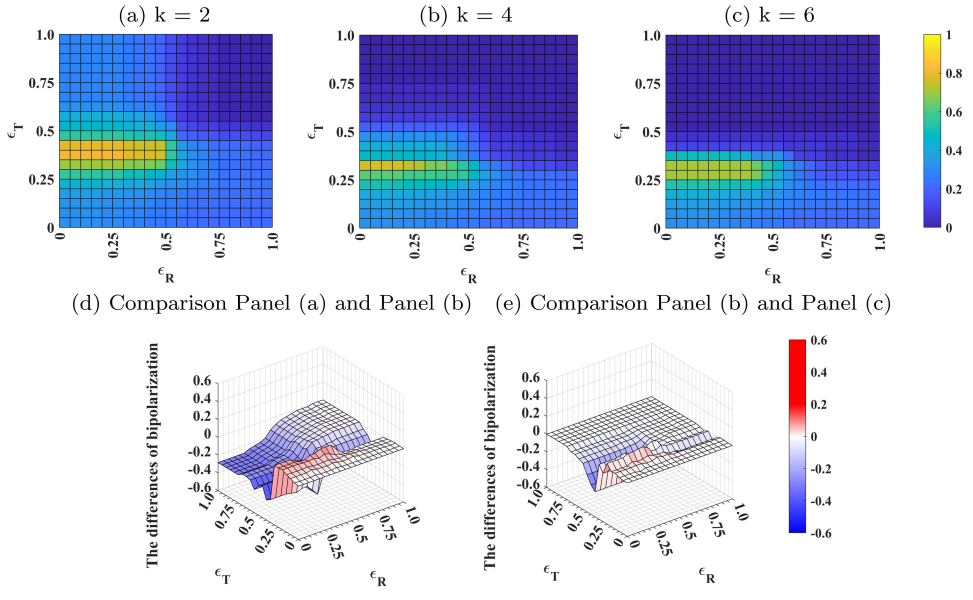


Fig. 8. Heat map of average degree of bipolarization in equilibrium for different trust and repulsion thresholds. (a) Negative links account for 10% of the total number of links ($k = 2$). (b) Negative links account for 5% of the total number of links ($k = 4$). (c) Negative links account for 3.3% of the total number of links ($k = 6$). (d) Heat map of the difference in the degree of bipolarization as the proportion of negative links varies from 10% to 5%. (e) Heat map of the difference in the degree of bipolarization as the proportion of negative links varies from 5% to 3.3%.

that runs ended in perfect consensus in larger parts of the parameter space when there were more positive network links.

However, a comparison of the specific cells of the three diagrams also supports the counter-intuitive effect. To visualize this, we plot in Panel (d) of Fig. 8 the difference between the average bipolarization values in Panels (a) and (b). Panel (e) compares bipolarization levels in Panels (b) and (c). Red areas indicate parts of the parameter space where we observed higher levels of bipolarization in the runs with a higher number of positive network links. This shows that the counter-intuitive effect shown in the stylized examples of Fig. 6 can be replicated for specific regions of the parameter space also with different initial opinion distributions. Nevertheless, Fig. 8 also shows that the counter-intuitive effect is small. This may simply be the result of the small population size of only 10 agents, a setting where a single random deviation from the engineered initial opinion distribution that generated the nonlinear effect in the first experiment can have a huge impact. Accordingly, we study bigger populations with more complex network structures in the following section.

3.2.2. *Network structure*

Our next robustness test was to increase the complexity of the network structure. We retained the main characteristic of the initial scenario in which a small share of

negative links induces a potential for repulsive influence to drive a population apart, whereas whether and to what extent this happens depends on the degree to which opinion dynamics lead the group to split along a more subtle "faultline" that separates different opinion clusters in the initial distribution. To assess the robustness of our main findings under a more complex network structure, we used a general network generator to create networks exhibiting to a varying degree the properties of a small-world network structure [38, 39]. While still an abstraction of the more complex features of real online social networks, small-world networks combine two features observed in many real-world social networks, a high degree of local clustering of ties and a short average path length, reflecting that closely-knit local communities are typically connected via "long-range" [3, 12] or "weak" ties [15], which preserve the overall connectivity of a population.

More specifically, the networks we constructed contain N nodes, where each node is connected with k links to its nearest neighbor. Each link in the network is reconnected randomly with probability p. That is, one endpoint of the link remains unchanged and the other endpoint is taken to another endpoint chosen randomly in the network. A completely regular network is induced with $p = 0$, while $p = 1$ corresponds to a completely random network. There can be at most one edge between any two different nodes, and nodes cannot be connected to themselves.

We created networks with small-world properties, setting the network size to $N = 100$ and the rewiring probability to $p = 0.05$. In this structure, we do not explicitly divide the group into two opposing groups. Positions of the negative links in the network are randomly distributed. As before, initial opinions of the nodes in the network are uniformly distributed between 0 and 1 with a tendency towards formation of two opinion clusters. The upper half of the nodes in the ring network prior to rewiring receives opinions between 0 and 0.5, and the lower half between 0.5 and 1. Each simulation experiment was run for 1000 time steps to ensure that the network reached a steady state when outcome measures were observed.

To test whether we could replicate the first set of counter-intuitive findings, we started with an experiment in which the average degree of each node is set to $k = 6$ and the ratio of negative links to total links in the network was 6.7%. Figure 9 shows three representative opinion dynamics given a repulsion threshold of $\epsilon_R = 0.5$ and trust thresholds of $\epsilon_T = 0.1, \epsilon_T = 0.3$ and $\epsilon_T = 0.5$, respectively. The color of the lines represents agents' initial opinion. In all three simulations, the same randomly generated initial opinion distribution and the same randomly generated network were used.

Panel (a) of Fig. 9 shows that the smallest trust threshold ϵ_T generates dynamics leading to the formation of multiple subgroups. Due to the moderate repulsion threshold, only a small number of individuals who are already inclined to extreme opinions and are connected by negative links adopt extreme opinions over time and manage to pull some similar friends with them. The equilibrium in this condition is characterized by a number of coexisting opinion clusters, yielding a medium level of bipolarization. Panel (b) of Fig. 9, in contrast, shows that the increase in the trust

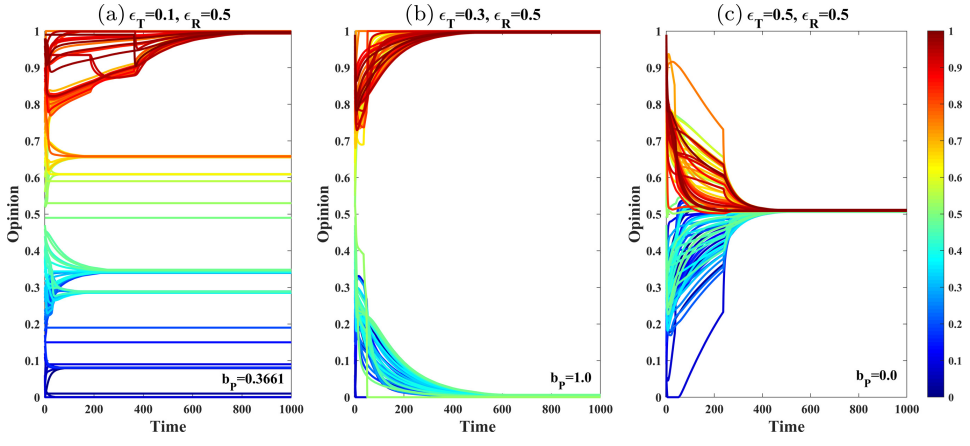


Fig. 9. Typical evolution of opinions in a small-world network with different trust thresholds. The repulsion threshold is fixed at $\epsilon_R = 0.5$. The trust threshold is set to $\epsilon_T = 0.1$ (a), $\epsilon_T = 0.3$ (b) and $\epsilon_T = 0.5$ (c), respectively.

threshold resulted in very strong opinion bipolarization. Furthermore, when we increased the trust threshold to 0.5 (Panel (c)), the run ended in consensus. The bipolarization measure b_p displayed in the figure shows that across conditions, the degree of bipolarization tends to increase from $\epsilon_T = 0.1$ to $\epsilon_T = 0.3$, and then decrease as the trust threshold increases to $\epsilon_T = 0.5$. This replicates the pattern we found for the simple ring network under otherwise the same conditions. Thus, the nonmonotonous effect of trust threshold on opinion bipolarization can also be generated on small-world networks.

To test the second counter-intuitive effect, we constructed three kinds of small-world network structures with different network node degrees. The average degree of nodes in these three types of network structures is $k = 2$, $k = 4$ and $k = 6$, respectively. Each network contained 100 nodes. The rewiring probability was set to $p = 0.05$. In order to analyze the effect of the increase in the proportion of positive links on opinion bipolarization, we fixed the number of negative links in the network, which implies a share of negative links of $P_n = 20\%$ under $k = 2$, $P_n = 10\%$ under $k = 4$, and $P_n = 6.7\%$ under $k = 6$. Figure 10 shows for three selected runs the evolution of node opinions in the network under different average node degrees. The box plots in Fig. 11 visualize the distribution of the bipolarization measure under the three different network structures at the point where the system reached equilibrium, based on 20 simulations per condition. In all conditions, we assumed $\epsilon_T = 0.3$ and $\epsilon_R = 0.5$.

From Fig. 10(a), it can be seen that most nodes' opinions remain unchanged and form isolated opinion clusters, so that opinion bipolarization does not differ much from the initial level when the network reaches the steady state. In the box plot, this is reflected by a small variation of the bipolarization measure around its approximate

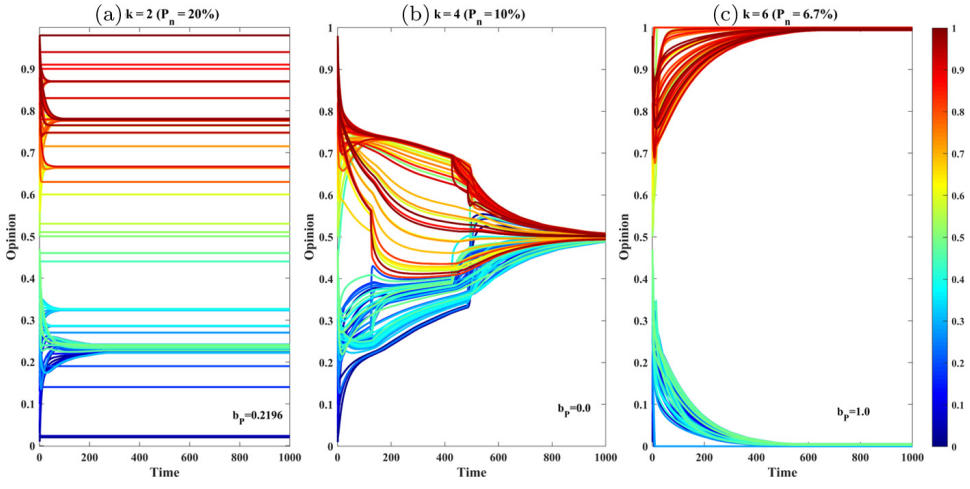


Fig. 10. Three representative graphs of the evolution of node opinions in the network under different average node degrees. (a) $k = 2$ ($P_n = 20\%$). (b) $k = 4$ ($P_n = 10\%$). (c) $k = 6$ ($P_n = 6.7\%$).

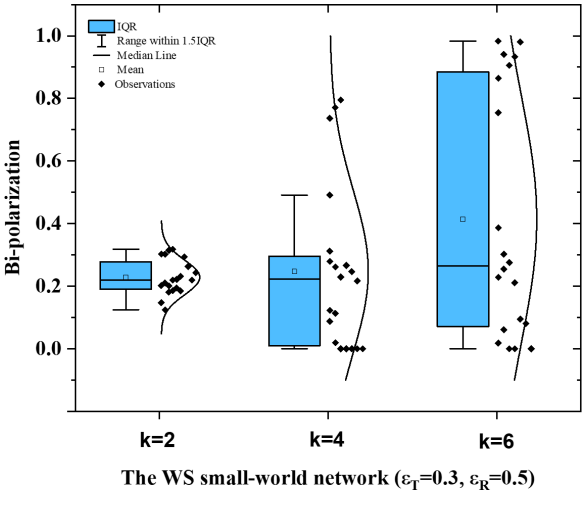


Fig. 11. Box plots of the distribution of the bipolarization measure in equilibrium, based on 20 simulation runs. The trust threshold is set to $\epsilon_T = 0.3$, repulsion threshold to $\epsilon_R = 0.5$. Three different small worlds are shown with (a) $k = 2$, (b) $k = 4$ and (c) $k = 6$, so that negative edges account for 20%, 10% and 6.7% of the total number of edges, respectively.

initial mean value of about 0.2. The reason for this result is the small number of connected dyads at $k = 2$. In this small number of dyads, the conditions for assimilative or repulsive influence are met in only a very few cases, which then have little impact on the overall opinion dynamics. When the average degree of nodes in Panel (b) reaches $k = 4$, the connectivity of the network increases. Figure 10(b) shows a case where this results in gradual convergence of all opinions towards consensus,

starting from assimilative influence between initially relatively similar individuals. However, the box plot in Fig. 11(b) demonstrates that there is now more variation in final outcomes. While a considerable proportion of runs indeed ends up in consensus (bipolarization measure of zero), some runs generate a high level of bipolarization and many fall in between those extremes. This shows that it depends sensitively on the initial opinion distribution and network structure whether repulsive influence is triggered for a sufficiently large share of nodes to induce higher levels of bipolarization or whether a number of nodes remain isolated from outside influences, producing intermediate levels of bipolarization. Finally, when the average degree of nodes increases to $k = 6$, outcomes with a high level of bipolarization become more likely. Figure 10(c) shows a typical run from this category, when $k = 6$ in Fig. 11 demonstrates that now most runs end up either close to maximal bipolarization or close to consensus, with some fraction still remaining in a state in between those extremes. The reason is that with higher connectivity of the network the repulsion mechanism is now more likely triggered for nodes at the interface of the two initial subgroups. This happens, because the larger connectivity within subgroups promotes a faster divergence between them at the interfaces, due to assimilative influence towards different local means. The combination of these two processes leads to an increase in the probability of network bipolarization. At the same time, if in this structure the nodes at both ends of a negative edges do not sufficiently diverge to trigger the repulsion mechanism, the larger connectivity can also promote consensus due to the larger share of positive edges. This result differs from the results of the highly stylized network of Fig. 6, which may be due to the difference in the size of the network and the initial opinion distribution, resulting in different threshold conditions required to form different situations. To further test the robustness of the changes in the distribution of bipolarization shown in Fig. 11, we replicated the experiment with a larger number of simulation. Results are shown in Fig. A.1.

Next, to test whether these nonmonotonous effects of changing network structure hold up also for different combinations of trust and repulsion thresholds, we executed multiple simulations with the three different network structures used above, varying for all three structures both ϵ_T and ϵ_R systematically from zero to one in steps of 0.05. We conducted 20 simulation runs for each condition. The rules for setting up the network structure in this experiment were the same as for the experiment shown in Fig. 11. As before, initial opinions of the nodes and the small-world network structure differ across runs in the same condition due to initial randomization.

Figure 12 shows for all three network structures a heat map of the average degree of bipolarization broken down by ϵ_T and ϵ_R values. Bipolarization is measured when opinion evolution reached a steady state in every condition. Since Fig. 12 reports only the average degree of bipolarization across multiple runs, we report in Fig. A.2, histograms of the bipolarization measure for three central parameter combinations.

From the three upper subplots in Fig. 12 we find that in small-world network structures with different proportions of negative edges, an increase in the trust threshold starting from $\epsilon_T = 0$ tends to first increase then decrease the degree of

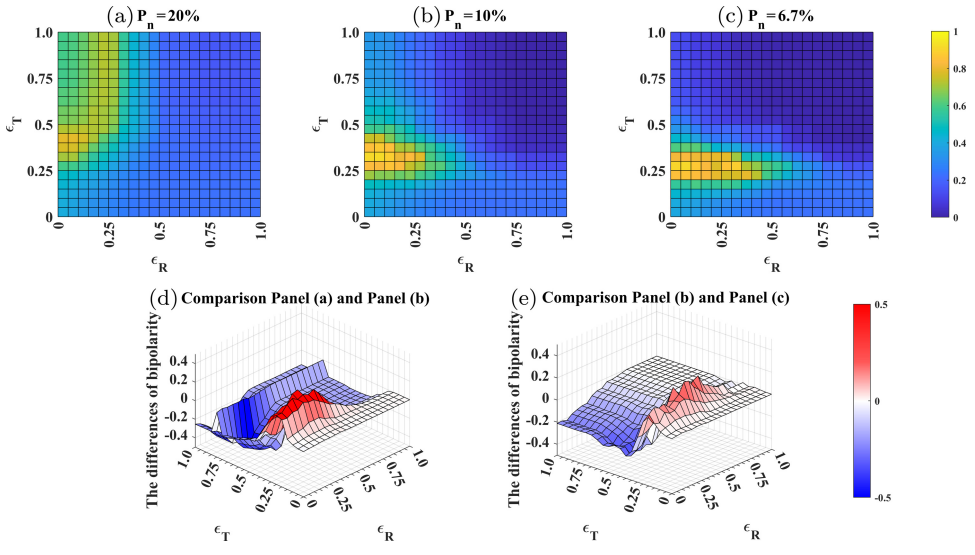


Fig. 12. Heat map of the degree of opinion bipolarization in the small-world networks at steady state, averaged over 20 simulations under different trust and repulsion thresholds. (a) Negative edges account for 20% of the total number of edges ($P_p = 1 - P_n(20\%)$). (b) Negative edges account for 10% of the total number of edges ($P_p = 1 - P_n(10\%)$). (c) Negative edges account for 6.7% of the total number of edges ($P_p = 1 - P_n(6.7\%)$). (d) Heat map of the difference in the degree of bipolarization between Panel (b) and Panel (a) (20% negative links versus 10%). (e) Heat map of the difference in the degree of bipolarization between Panel (c) and Panel (b) 10% to 6.7%.

bipolarization when the repulsion threshold is sufficiently low. This reflects our earlier finding that more assimilative influence can promote bipolarization, but also shows how this effect is limited to regions of the parameter space with relatively low values of trust and repulsion thresholds across all three network structures.

Especially in networks with weak connectivity, an increase in the trust threshold in this region of the parameter space causes nodes with opinions close to those of extreme neighbors to converge towards them, which explains why we find an increase in the degree of bipolarization as ϵ_T increases. At the same time, in networks with weak connectivity consensus is unlikely to form even when the repulsion threshold is very high (with a very small effect of negative influence).

Similarly, for the condition $k = 4$, we find that when the repulsion threshold is below 0.5, the increase in the number of positive edges promotes bipolarization within the range of trust threshold values between 0.25 and 0.55. And, for the condition $k = 6$, the further increase in the share of positive edges extends the range of values for the repulsion threshold for which we find a positive effect of higher trust thresholds on bipolarization in the range between $\epsilon_T = 0.2$ and $\epsilon_T = 0.35$. For higher values of ϵ_T bipolarization drops to approximately zero for both $k = 4$ and $k = 6$ and sufficiently high repulsion threshold, reflecting that in the two networks with higher connectivity consensus prevails when the conditions for assimilative influence become less restrictive and those for repulsive influence more restrictive.

The comparison of the bipolarization measures between different network structures (Panels (d) and (e)) highlights that consistently with earlier findings, the increase in the proportion of positive edges can promote bipolarization in large regions of the parameter space. Especially in the region of the ϵ_R–ϵ_T plane where the repulsion threshold is less than about 0.5 and the trust threshold is between approximately 0.2 and 0.35, it can be seen that bipolarization increases as the share of negative edges falls from $P_n = 20\%$ to $P_n = 10\%$, and then increase even further between $P_n = 10\%$ to $P_n = 6.7\%$. At the same time, in some areas the nonmonotonic effects of positive edges can be found, as for example for $\epsilon_T = 0.4$ and $\epsilon_R = 0.1$, where bipolarization slightly increases between Panels (a) and (b), then drops between Panels (b) and (c).

4. Conclusions

In the public and scholarly debate about the effects of online social networks on opinion bipolarization, there are two opposing arguments. On the one hand, there is growing concern that online social networks could contribute to the bipolarization of opinions. The personalization of the information users consume on these platforms might create filter bubbles shielding users from content that challenges their views and preventing assimilative social influence between users with opposite political opinions. On the other hand, this warning has been challenged by the observation of repulsive influence, the counter-part of assimilative social influence. It has been observed that exposing users to content promoting opposing political views can generate opinion shifts away from the source of influence. When there is assimilative social influence, it has been argued, decreasing exposure to content with opposing opinions would prevent opinion bipolarization rather than intensify it. In a nutshell, we argued here that both arguments fail to capture the complexity of the interplay of assimilative and repulsive social influence in networks. We demonstrated with an agent-based model that the conjunction of assimilative and repulsive social influence can generate counter-intuitive opinion dynamics, including the possibility that more assimilative influence may actually promote bipolarization in a population. This suggests that predicting whether personalization technology contributes to opinion bipolarization or not is not trivial and requires a formal analysis.

To demonstrate the counter-intuitive consequences of assimilative and repulsive social influence, we studied highly stylized examples, assuming very small populations with a simple network structure and a deterministic opinion updating schedule. Despite this simplicity, we showed that — counter-intuition — giving assimilative social influence more room can sometimes generate more and not less opinion bipolarization. While assimilative influence is a strong force generating global opinion convergence in populations, it can also contribute to local opinion convergence. That is, assimilative social influence can contribute to the formation of homogeneous subgroups. While individuals within subgroups hold similar opinions, opinion differences between subgroups can be intensified to a degree that repulsive influence

between groups is activated. As a consequence, more assimilative social influence can generate opinion bipolarization.

We also demonstrated that this counter-intuitive effect does not only emerge under the highly stylized conditions of our simple example networks. When the network is characterized by a subtle "hidden" faultline, opinion differences between subgroups can be intensified by assimilative social influence and create the breeding ground for bipolarization.

What conclusions can we draw about the contribution of the personalization of online social networks to processes of opinion bipolarization? The central insight from our analyses is that based on the empirical research on social-influence processes alone, one cannot make inferences about whether or not personalization is responsible for growing opinion bipolarization. Since even simple, highly stylized networks generate counter-intuitive dynamics it is not possible to draw conclusions about the effects of social influence in systems as complicated as real online social networks. Online social networks are complex systems. Deriving conclusions about the effects of micro-mechanisms such as social influence requires a rigorous, formal analysis. Obviously, one could counter these conclusions with the argument that we engineered the model and the networks in a way that it generates counter-intuitive dynamics. This is true and it may be possible to construct a model with assimilative and repulsive influence that may not generate the same dynamics or do so only under different conditions. It is true that often even small changes in the assumptions of a model can generate very different model predictions. This observation, however, should not lead one to the conclusion that a rigorous theoretical analysis is not necessary. In contrast, if small things matter, a proper theoretical analysis is even more important, as human intuition tends to overlook seemingly small effects. In other words, the take-home message of our modeling exercise is not that assimilative and repulsive influence have counter-intuitive macro-consequences, but that they may have them. The conclusion to draw from this insight is that one should conduct the necessary modeling research before making inferences about macro-outcomes.

Hence, while we cannot draw conclusions about whether or not online social networks are responsible for opinion polarization, we can derive an important recommendation for future research. On the one hand, we applaud empirical research into the micro-processes affecting opinion adjustments resulting from online interaction. Empirical methods applied in the growing field of computational social science are vital to understanding the effects of digital communication on opinion dynamics. On the other hand, the relationship between these micro-processes and their macro-consequences may be more complex than intuition suggests. As a consequence, drawing conclusions about the macro-effects of the micro-processes active on online social networks requires a rigorous formal analysis. Examples of how empirical observations of micro-level processes can be translated into macro-level predictions can be found in the work by Del Vicario *et al.* [7] and the work by Keijzer *et al.* [24].

While this paper applied models of opinion dynamics to demonstrate that intuition often fails, these models can also be used to rigorously identify conditions under which bipolarization emerges. We have shown, for instance, that bipolarization can rise despite stronger assimilative influence when assimilative influence generates local opinion convergence, which later activates repulsive influence. Accordingly, a testable hypothesis about the conditions of bipolarization would be that bipolarization is more likely to emerge when there is a potential for local opinion convergence and when there are negative relationships between the local subgroups. Future modeling work is needed to study this insight in larger networks.

The call for more formal modeling does not only address the debate about online social networks but research on bipolarization in general. As discussed at the outset, the analyses presented in this paper also contribute to the debate about affective polarization or "sectarianism" [10, 20]. Yet, while we focused here on the effect of a fixed structure of positive and negative intergroup relations, we believe that future work should model valence of intergroup relations as an endogenous phenomenon that can change — albeit perhaps at a different time scale — in response to agreement or disagreement on substantive topics in the public debate. Empirical research, for instance, documents rising affective polarization in the US in terms of increasingly negative evaluations of individuals holding different political opinions [2, 10]. However, as we showed, the prevalent intuition that growing negative affect between members of different political camps fosters repulsive influence [9, 14, 17] and, as a consequence, opinion bipolarization, turned out to be incomplete. This can also have implications for the debate about bursting filter bubbles by creating more contact between users of online social networks who dislike each other. While some argue that this may foster rather than mitigate opinion polarization [1] by allowing for more repulsive influence, we showed that also the opposite effect is possible. The stylized examples shown in Fig. 6 revealed that decreasing affective polarization by adding more positive social relationships to a network can lead to more and not less opinion bipolarization.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China under Grant 71871042 (to HX), the Humanities and Social Science Project of the Ministry of Education of China Grant 18YJA630118 (to HX). SL gratefully acknowledges the financial support provided by China Scholarship Council (CSC) through the program public postgraduates of national construction high-level universities (202006060168). MM acknowledges support by the InfoXpand project (031L0300B) funded by the Federal Ministry of Education and Research of Germany and support by the Netherlands Institute for Advanced Studies in the Humanities and Social Sciences (NIAS) and the Lorentz Center for funding the Lorentz-NIAS Theme Group "Social Media for Digital Democracy" that informed this work. AF acknowledges financial support by the Netherlands Organization for Scientific Research (NWO) under the 2018 ORA grant ToRealSim (464.18.112).

Appendix A

A.1. *Detailed analyses of circle network with random initial opinions*

Figure 7 demonstrated the robustness of the counter-intuitive effect to changes in the initial opinion distribution. Since we observed a high variation of bipolarization under the condition with $\epsilon_T = 0.3$, we report in Fig. A.1 a histogram of a simulation experiment with 500 rather than only 20 replications. Confirming the findings reported above and providing evidence for the robustness of the counter-intuitive effect of increasing the openness to assimilative influence, the histogram shows that the majority of the simulations ended in a state of perfect bipolarization.

A.2. *Detailed analyses of the two counter-intuitive effects in small-world networks*

Since the heat maps of Fig. 12 report only the average degree of bipolarization, we inform in Fig. A.2 about the distribution of bipolarization in 500 simulation runs conducted in three conditions. For all runs, we assumed a small-world network with a degree of $k = 6$ and a rewiring probability of $p = 0.05$, which corresponds to Panel (c) in Fig. 12. The repulsion threshold was set to $\epsilon_T = 0.5$ in all runs. The trust threshold ϵ_R was varied between 0.1 and 0.5 in steps of 0.2, in order to show the distribution of bipolarization in the three most interesting phases. Under $\epsilon_R = 0.1$, all runs ended in a state of fragmentation with multiple opinion clusters with

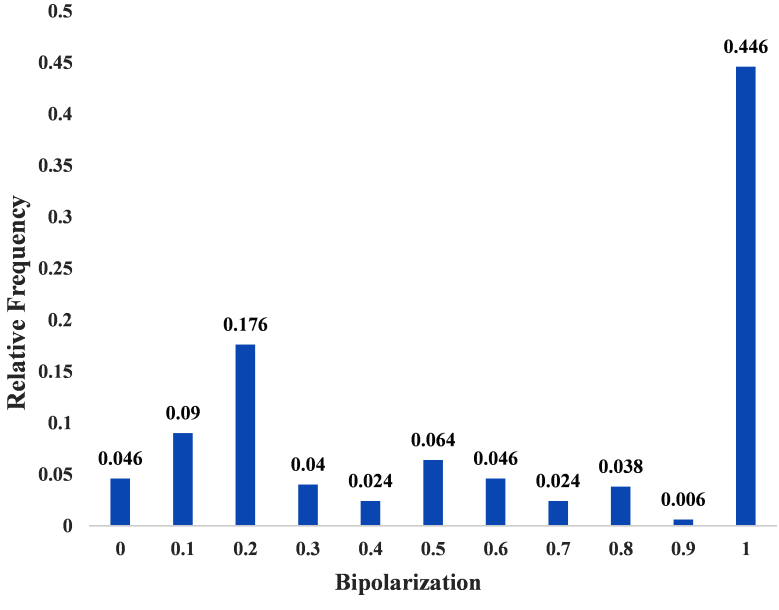


Fig. A.1. Histogram depicting the distribution of bipolarization in 500 runs with the stylized example shown in Fig. 3 but different initial opinions ($\epsilon_T = 0.3$, $\epsilon_R = 0.5$).

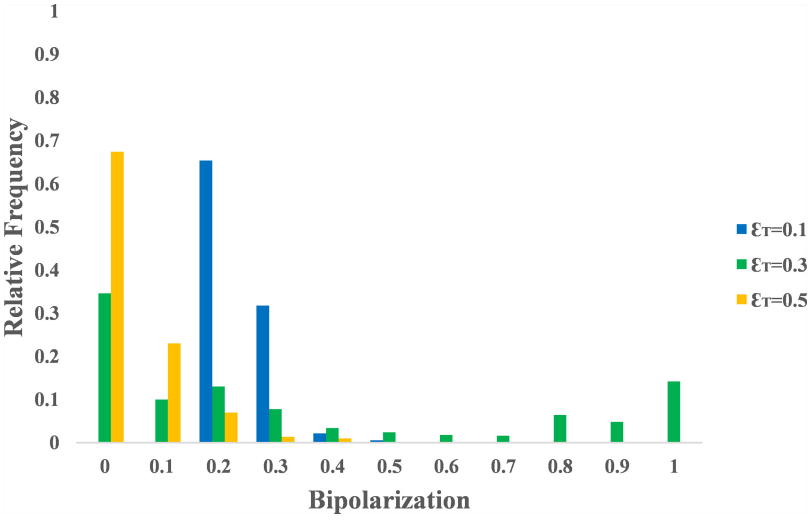


Fig. A.2. Histograms of bipolarization in 500 independent runs under different values of the trust threshold. In all runs, we assumed a small-world network with $k = 6$ and $p = 0.05$, and set the repulsion threshold to $\epsilon_R = 0.5$. The trust threshold (ϵ_T) was set to 0.1, 0.3 and 0.5, respectively.

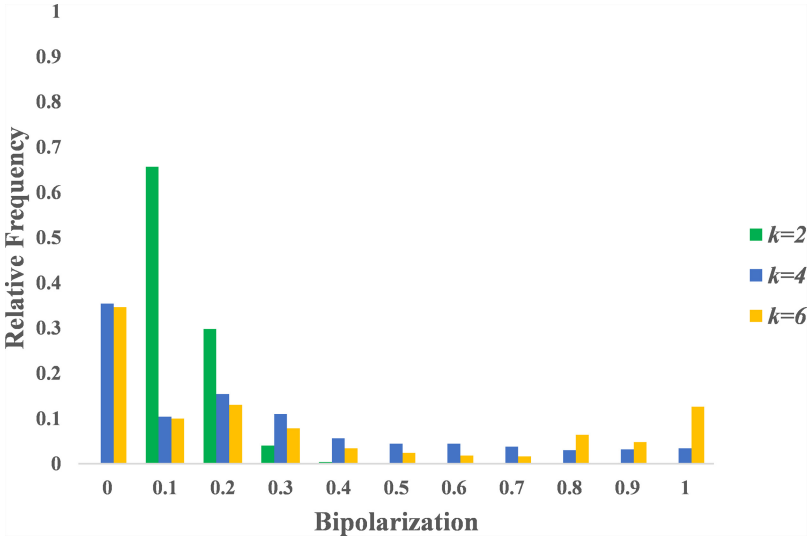


Fig. A.3. Histograms of bipolarization in 500 independent runs under different numbers of positive network relationships. In all runs, we assumed a small-world network with $p = 0.05$, set the repulsion threshold to $\epsilon_R = 0.5$ and the trust threshold to $\epsilon_T = 0.3$. The average degree k of positive ties was varied from 2 to 6 in steps of 2.

nonextreme opinions. Under $\epsilon_R = 0.2$, the intermediate phase, about a third of the runs reached consensus, but another 15% of the runs ended in perfect bipolarization into two opposing groups, which replicates the counter-intuitive effect of increasing assimilative social influence. When ϵ_R is increased even more, runs tend to end in consensus or opinion distributions with a low variance.

Figure A.3 shows the same analyses for the second counter-intuitive effect. To this end, we varied the average degree k in the network and kept all other parameters constant. That is, we set the repulsion threshold to $\epsilon_R = 0.5$ and the trust threshold to $\epsilon_T = 0.3$. The green bars show the histogram of bipolarization when the degree was set to $k = 2$, which is the same condition as shown by the blue bars in Fig. A.2. This histogram shows that all runs ended in a state of fragmentation. When k is increased to four (see the blue bars in Fig. A.3), most runs end in a state of perfect opinion consensus. There is also a small share of the runs characterized by very high bipolarization. Strikingly, very high values of bipolarization are more likely when the degree is increased further to $k = 6$, as the yellow bars reveal. This replicates the non-monotonic effect of the share of positive network ties on bipolarization at steady state.

References

[1] Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., Lee, J., Mann, M., Merhout, F. and Volfovsky, A., Exposure to opposing views on social media can increase political polarization, *Proc. Natl. Acad. Sci.* **115** (2018) 9216–9221.

[2] Boxell, L., Gentzkow, M. and Shapiro, J. M., Greater internet use is not associated with faster growth in political polarization among us demographic groups, *Proc. Natl. Acad. Sci.* **114** (2017) 10612–10617.

[3] Centola, D. and Macy, M., Complex contagions and the weakness of long ties, *Am. J. Sociol.* **113** (2007) 702–734.

[4] Dandekar, P., Goel, A. and Lee, D. T., Biased assimilation, homophily, and the dynamics of polarization, *Proc. Natl. Acad. Sci.* **110** (2013) 5791–5796.

[5] Deffuant, G., Bertazzi, I. and Huet, S., The dark side of gossips: Hints from a simple opinion dynamics model, *Adv. Complex Syst.* **21** (2018) 1850021.

[6] Deffuant, G., Neau, D., Amblard, F. and Weisbuch, G., Mixing beliefs among interacting agents, *Adv. Complex Syst.* **3** (2000) 87–98.

[7] Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E. and Quattrociocchi, W., The spreading of misinformation online, *Proc. Natl. Acad. Sci.* **113** (2016) 554–559.

[8] Del Vicario, M., Zollo, F., Caldarelli, G., Scala, A. and Quattrociocchi, W., Mapping social dynamics on Facebook: The Brexit debate, *Soc. Netw.* **50** (2017) 6–16.

[9] Festinger, L., *A Theory of Cognitive Dissonance*, Vol. 2 (Stanford University Press, 1957).

[10] Finkel, E. J. *et al.*, Political sectarianism in America, *Science* **370** (2020) 533–536.

[11] Flache, A., About renegades and outgroup haters: Modeling the link between social influence and intergroup attitudes, *Adv. Complex Syst.* **21** (2018) 1850017.

[12] Flache, A. and Macy, M. W., Small worlds and cultural polarization, *J. Math. Sociol.* **35** (2011) 146–176.

[13] Flache, A. and Mäs, M., Why do faultlines matter? A computational model of how strong demographic faultlines undermine team cohesion, *Simul. Model. Pract. Theory* **16** (2008) 175–191.

[14] Flache, A., Mäs, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S. and Lorenz, J., Models of social influence: Towards the next frontiers, *J. Artif. Soc. Soc. Simul.* **20** (2017).
[15] Granovetter, M. S., The strength of weak ties, *Am. J. Sociol.* **78** (1973) 1360–1380.
[16] Hegselmann, R., 9 polarization and radicalization in the bounded confidence model: A computer-aided speculation, in *Advances in the Sociology of Trust and Cooperation: Theory, Experiment, and Field Studies* (De Gruyter, 2020), pp. 197–226.
[17] Heider, F., Attitudes and cognitive organization, *J. Psychol.* **21** (1946) 107–112.
[18] Huet, S. and Deffuant, G., Openness leads to opinion stability and narrowness to volatility, *Adv. Complex Syst.* **13** (2010) 405–423.
[19] Huet, S., Deffuant, G. and Jager, W., A rejection mechanism in 2D bounded confidence provides more conformity, *Adv. Complex Syst.* **11** (2008) 529–549.
[20] Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N. and Westwood, S. J., The origins and consequences of affective polarization in the United States, *Annu. Rev. Polit. Sci.* **22** (2019) 129–146.
[21] Jager, W. and Amblard, F., Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change, *Comput. Math. Organ. Theory* **10** (2005) 295–303.
[22] Keijzer, M. A. and Mäs, M., The strength of weak bots, *Online Soc. Netw. Media* **21** (2021) 100106.
[23] Keijzer, M. A. and Mäs, M., The complex link between filter bubbles and opinion polarization, *Data Sci.* **5** (2022) 139–166.
[24] Keijzer, M. A., Mäs, M. and Flache, A., Polarization on social media: Micro-level evidence and macro-level implications, in *Opinion Dynamics in Online Social Media*, Chap. 2 (University of Groningen, 2022).
[25] Kozitsin, I. V., Opinion dynamics of online social network users: A micro-level analysis, *J. Math. Sociol.* (2021) 1–41.
[26] Levin, S. A., Milner, H. V. and Perrings, C., The dynamics of political polarization, *Proc. Natl. Acad. Sci.* **118** (2021) e2116950118.
[27] Liu, C. C. and Srivastava, S. B., Pulling closer and moving apart: Interaction, identity, and influence in the U.S. senate, 1973 to 2009, *Am. Sociol. Rev.* **80** (2015) 192–217.
[28] Macy, M. W., Kitts, J. A., Flache, A. and Benard, S., Polarization in dynamic networks: A Hopfield model of emergent structure, in *Dynamic Social Network Modeling and Analysis* (National Academies Press, 2003).
[29] Macy, M. W. and Willer, R., From factors to actors: Computational sociology and agent-based modeling, *Annu. Rev. Sociol.* **28** (2002) 143–166.
[30] Mäs, M., Analytical sociology and complexity research, in *Research Handbook on Analytical Sociology* (Edward Elgar Publishing, 2021), pp. 100–118.
[31] Mäs, M. and Flache, A., Differentiation without distancing. Explaining bi-polarization of opinions without negative influence, *PLoS One* **8** (2013) e74516.
[32] Obama, B., President Obama's farewell address: Full video and text, The New York Times (2017).
[33] Pariser, E., *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think* (Penguin, 2011).
[34] Schweighofer, S., Schweitzer, F. and Garcia, D., A weighted balance model of opinion hyperpolarization, *J. Artif. Soc. Soc. Simul.* **23** (2020) 5.
[35] Schelling, T. C., *Micromotives and Macrobehavior* (WW Norton & Company, 2006).
[36] Steinmeier, F.-W., 2018 Christmas message (2018).
[37] Takács, K., Flache, A. and Mäs, M., Discrepancy and disliking do not induce negative opinion shifts, *PLoS One* **11** (2016) e0157948.

[38] Watts, D. J., Networks, dynamics, and the small-world phenomenon, *Am. J. Sociol.* **105** (1999) 493–527.
[39] Watts, D. J. and Strogatz, S. H., Collective dynamics of 'small-world' networks, *Nature* **393** (1998) 440–442.
[40] Zuckerberg, M., Building global community, Facebook (2017).