Opinion Homogenization and Polarization - Three Sampling Models

Elizaveta Konovalova

University of Warwick, Warwick Business School

Gaël Le Mens

Universitat Pompeu Fabra, Department of Economics and Business

Barcelona School of Economics

UPF-Barcelona School of Management

Author Note

G. Le Mens benefited from financial support from grant PID2019-105249GB-I00/AEI/10.13039/501100011033 from the Spanish from the Spanish Ministerio de Ciencia, Innovacion y Universidades (MCIU) and the Agencia Estatal de Investigacion (AEI), ERC Consolidator Grant #772268 from the European Commission and BBVA Foundation Grant G999088Q. Emails: elizaveta.konovalova@wbs.ac.uk & gael.le-mens@upf.edu.

Abstract

We describe three sampling models that aim to cast light on how some design features of social media platforms systematically affect judgments of their users. We specify the micro-mechanisms of belief formation and interactions and explore their macro implications such as opinion polarization. Each model focuses on a specific aspect platform-mediated social interactions: how popularity creates additional exposure to contrarian arguments, how differences in popularity make an agent more likely to hear particularly persuasive arguments in support of popular options, and how opinions in favor of popular options are reinforced through social feedback. We show that these mechanisms lead to self-reinforcing dynamics that can result in local opinion homogenization and between-group polarization. Unlike non-sampling-based approaches, our focus does not lie in peculiarities of information processing such as motivated cognition but instead emphasizes how structural features of the learning environment contribute to opinion homogenization and polarization.

Introduction

The polarization of opinions has been described as a "challenge to democratic debate" (EU Commissioner Vera Jourova) and is frequently seen as an important problem that needs to urgently be addressed in order to preserve social harmony. A number of commentators and politicians have attributed opinion polarization to the abundance of 'fake news' that spread on social media and via more traditional channels such as cable television and the press. The design of social media platforms, and how these shape the information to which people have access, have also been pointed out as culprits. For example, Sunstein (2018) argued that because social media platforms more strongly connect like-minded people than people with different opinions, this facilitates the emergence of opinion clusters. This clustered structure would, in turn, affect the kind of information to which users are exposed, contributing to a self-reinforcing dynamic. Others have proposed that this self-reinforcing dynamic could be strengthened by the algorithms that control the content shown to users. These would create 'filter bubbles' (Pariser, 2011) in which people fail to be exposed to contrarian ideas and opinions. This could, in turn, contribute to an increase in the popularity of extreme opinions, contributing to polarization on issues such as climate policy, environmental issues, social policy, colonial history, immigration, gender issues, or abortion rights.

These claims that social media contribute to opinion polarization fall within the scope of the sampling approach to human judgment because they focus on how social media shape the information to which their users have access. Yet, they do not clearly spell out the micro-mechanism of social influence at play and how such micro-mechanisms could contribute to a macro phenomenon such as between group opinion polarization. In this chapter, we aim to address this shortcoming by discussing three sampling

models that focus on distinct aspects of how social interactions happen on social media platforms: how popularity creates additional exposure to contrarian arguments, how differences in popularity make an agent more likely to hear particularly persuasive arguments in support of popular options, and how opinions in favor of popular options are reinforced through social feedback. The first two models pertain to how newsfeeds — a personalized flow of information that contains posts from other users or companies shown to users of social media as they login onto the platform— affect opinion dynamics. They assume that arguments for opinions popular among the network contacts of a user will be prevalent in their newsfeed and illustrate two mechanisms according to which this can contribute to social influence. The third model pertains to the effect of feedback in the form of 'likes,' 'favorites,' and 'retweets' — an integral part of social media that has been shown to provide a motivation for people to use such platforms (Eckles, Kizilcec, & Bakshy, 2016; Chen, Chen, & Agarwal, 2017). This model posits that people are more likely to express opinions that received positive feedback and that feedback is more likely to be provided by network contacts than by other users. Each sampling model leads to two jointly occurring phenomena: opinion homogenization between densely connected agents and opinion polarization between agents who are not connected or only indirectly connected.

This sampling perspective on homogenization and polarization builds on the idea of assimilative influence — a classical approach to polarization where the focal agent converges in their belief with others through social influence. A common approach used in discussions of models of assimilative influence assumes that people converge in their belief with others by simply adopting the average opinion of their peers (DeGroot, 1974; Friedkin, 1999; Latané, Nowak, & Liu, 1994). Even though this approach has been widely adopted in prior work on opinion dynamics in social networks, it is a kind of 'black box' regarding the micro-process of social influence because it does not specify the mechanism according to which the opinion of an agent would become aligned with that of the agents to which they are connected. Moreover, the lack of specificity of this perspective renders it similarly applicable to mechanisms that rely on motivated cognition, rational inferences, and sampling-based mechanisms. Here, we contribute to opening this 'black box' by specifying sampling-based mechanisms that lead to local opinion homogenization and between group polarization.

The most common explanation for local opinion homogenization in models of assimilative influence relies on motivated information processing as the result of the desire to belong and avoid punishment for deviating from the group norm: the social pressure applied by others motivates the person to interpret information such that the resulting opinion conforms to expectations of others (McGarty, Turner, Hogg, David, & Wetherell, 1992; Turner, Wetherell, & Hogg, 1989). Another explanation presumes that agents infer the 'quality' of a stance based on their perceived popularity. Research in economics has shown that it is sometimes rational to do so, and that such inferences can lead to opinion

homogenization through information cascades (Banerjee, 1992; Bikhchandani, Hirshleifer, & Welch, 1992).

In contrast, the sampling approach discussed in this chapter aims to specify conditions regarding how people sample information from their environments that are sufficient to produce homogenization or polarization even in the absence of motivated information processing or popularity-based inferences. The unifying theme of the three models is that they are particularly relevant to interactions on social media, but they are also applicable to other settings that satisfy their assumptions regarding how people sample information. Each model is simplistic, in the sense that it focuses on just one sampling mechanism while excluding other sampling mechanisms and 'switches off' motivated cognition and popularity-based inferences. In presenting these models, we do not claim that motivated cognition and popularity-based inferences are unimportant or that just one sampling mechanism is applicable in a given setting. Rather, we see these models as proofs-of-concept that these sampling mechanisms are each sufficient to produce homogenization and polarization. This implies, in particular, that if an analyst observes that opinions have become more homogeneous or more polarized in an empirical setting, several sampling mechanisms could have produced this phenomenon. Thus, explaining the reasons for the observed homogenization requires uncovering the specifics of the sampling process or processes (and/or information processing and inference mechanisms) at play.

Next, we first present an informal description of the three sampling models, then we discuss the relationship between social influence, local opinion homogenization, and between group polarization. We then turn to a formal analysis of the three models.

Three Sampling Models for Social Influence

Florian just arrived in Barcelona for his Erasmus exchange stay at a local university. He quickly realizes that most Catalans tend to have strong opinions about a particular issue: Catalan independence. Florian wonders what he should think about this issue: would an independent Catalonia be a good thing or a bad thing? As he talks to his classmates and reads the news articles his new friends post on Facebook and Twitter, Florian samples information about the issue and begins to form his own opinion. In this introductory section, we explain how such information sampling can lead Florian to favor independence when most people in his social circle are pro-independence. (Symmetric predictions would hold if most of his classmates were against independence). In what follows, we provide a verbal description of the three sampling-based social influence models and then move onto formal analyses of these models.

The 'Asymmetric Hot-Stove' model analyzes the dynamics of Florian's opinion as he samples pro-independence and anti-independence arguments in the content shared by his Facebook friends and those

he follows on Twitter or in conversations with his classmates.[1] When neither of the options is much more popular than the other, the Hot-Stove effect (see Chapter [Chapter number of Hot-Stove Effect chapter]) implies that Florian is likely to underestimate the value of one of the stances. This is because, if he samples unconvincing arguments in favor of this stance, he will avoid sampling further arguments about this stance and will fail to discover there is merit to this stance.

Now, consider the effect of the popularity of the pro-independence stance. Because the pro-independence stance is popular in Florian's social circle, the Hot-Stove effect will affect this stance less strongly than the anti-independence stance. Even if Florian starts to develop an anti-independence opinion and thus would rather avoid reading or hearing more pro-independence arguments, because most of his social circle is pro-independence, he keeps being exposed to the pro-independence perspective and thus obtains additional samples of information about it. By contrast, if Florian develops a pro-independence opinion, and comes to dislike the anti-independence stance, he will not be exposed to many additional arguments about this stance because it is unpopular in his social circle. This asymmetry in exposure to pro-independence and anti-independence arguments implies that Florian can be subject to a 'Hot-Stove' effect about the pro-independence stance but is unlikely to be subject to a Hot-Stove effect about the anti-independence stance (see the 'Hot-Stove' chapter). In other words, Florian is unlikely to underestimate his attraction for the pro-independence stance but underestimation of the anti-independence stance is possible. Overall, this makes Florian more likely to adopt an opinion in line with that of his social circle: pro-independence.

The 'Maximal Argument Strength' model also analyzes the dynamics of Florian's opinion as he samples pro-independence and anti-independence arguments in the content shared by his social circle.[2] In each period, Florian hears pro-independence and anti-independence arguments from his social circle and he is most influenced by the strongest argument he hears. And because the strongest arguments are likely to come from the larger sample of arguments, the strongest argument he hears is likely to support the more popular stance in his social circle. Because most of Florians' social circle prefers the pro-independence stance, this implies that Florian is likely to shift his opinion in favor of independence. By contrast to the 'Asymmetric Hot-Stove' model, the 'Maximal Argument Strength' model does not assume that Florian's current opinion affects the samples of information he will obtain (he does not engage in what is sometimes called 'active' sampling but instead 'passive' sampling [make a connection to some other chapters of the

[1] This model was initially introduced in Denrell and Le Mens (2007) and its implications for collective opinions were analyzed in Denrell and Le Mens (2017).

[2] This model is original to this chapter

book once we have seen them]). Here, the samples are entirely driven by the composition of Florian's social circle.

The 'Feedback Sampling' model examines what happens as Florian starts to express his opinions about Catalan independence by sharing content supporting his opinion on social media or by explicitly stating his opinion in conversations with friends and classmates.[3] When Florian expresses his opinion, he sometimes gets approval in the form of a 'like' or 'retweet' or a signal of approval in a conversation. At other times, he gets negative feedback, in the form of negative comments. Because most of Florian's classmates are pro-independence, he tends to get more positive feedback when he shares a pro-independence opinion than when he shares an anti-independence opinion. If Florian cares about such feedback, he will respond by shifting the position of his statements toward the pro-independence stance. This model differs from the other two models in terms of the unit of sampling. Whereas in the Asymmetric Hot-Stove model and the Maximal Argument Strength model, Florian forms opinions about the two stances based on arguments expressed by members of his social circle (he samples arguments), here, he voices his own opinions and arguments but samples feedback (that could be negative or positive) about the arguments he expresses.

In what follows, we discuss how these basic sampling-based social influence mechanisms can contribute to explaining opinion homogenization and polarization. The resulting patterns of opinions and preferences depend on the structure of the social network in which agents are embedded.

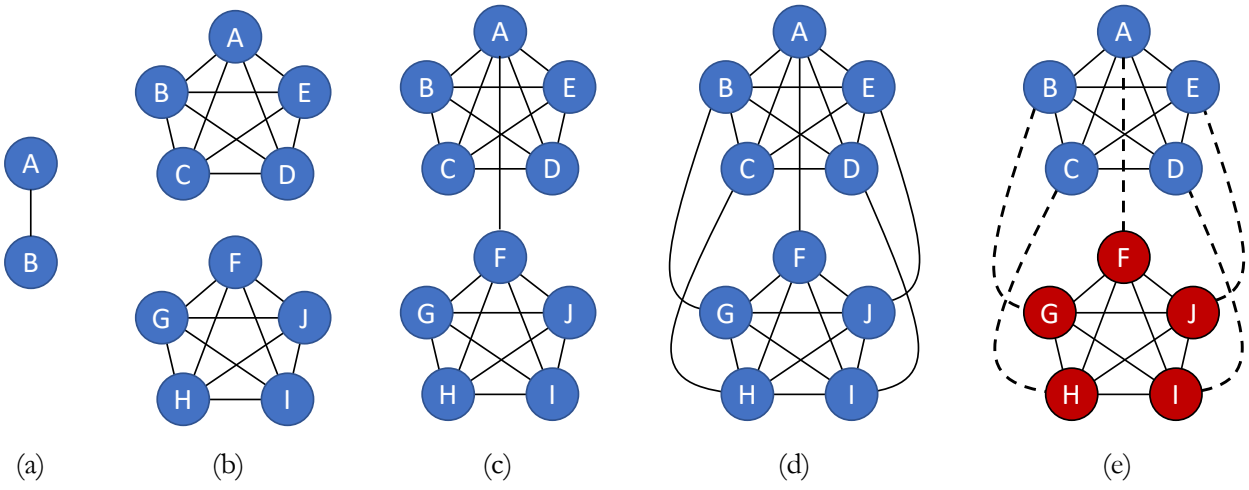

Figure 1: Social network structures analyzed in the simulations.

[3] This model builds on and on-going project by the authors of this chapter and Nikolas Schöll.

From Social Influence to Opinion Polarization

In all three models, the samples obtained by Florian are subject to randomness: In the 'Asymmetric Hot-Stove,' what is random is the strength of a sampled argument. In the 'Maximal Argument Strength', what is random is the strength of the argument of maximal strength among the set of sampled arguments. In the 'Feedback Sampling' model, what is random is the valence and strength of the feedback provided by Florian's social circle. The randomness at the core of each model implies that the same set of initial conditions can lead to different resulting patterns. For example, even if Florian's social circle favors independence, Florian might become anti-independence. Yet, all three models predict that Florian is more likely to adopt an opinion aligned with that of his social circles than an opposed opinion.

A more formal and general rendition of this claim is that the three sampling models imply that opinions become correlated. In other words, agents who are close to each other in the social network tend to come to prefer the same options even when initial evaluations are independent of each other and there is no clear initial majority. The mechanisms we discuss in this chapter thus explain the *emergence* of opinion homogeneity and preference convergence in densely connected networks. To unpack this process, after providing a formal description of the models, we first explain how they can lead to opinion homogenization and preference convergence in a tiny network of two agents (Fig. 1a) who update their opinions in each period.

We then turn to polarization. We call 'polarization' the *divergence* of opinions and preference between-groups of people or distinct parts of a social network.[4] We illustrate how the three social influence models can produce polarization in networks with 'structural holes' – agents who are not connected to each other (Burt, 2009). We first consider a simple setting in which agents belong to two independent (i.e., disconnected) groups and then settings in which the groups are more or less densely connected to each other. We consider two cases: one case in which group identity is irrelevant to the nature of the interaction between agents (Figs. 1b,c,d), and one case in which it is relevant (Fig. 1e, see Iyengar, Sood, and Lelkes (2012)).

The 'Asymmetric Hot-Stove' Model

[4] We do not use the word 'polarization' how it was used in the program of research on 'attitude polarization' in social psychology (Burnstein and Vinokur (1977); Moscovici and Zavalloni (1969); Nowak, Szamrej, and Latané (1990); for reviews, see Isenberg (1986) and Myers and Lamm (1976)). This research program focused on what we call *local opinion homogenization* rather than polarization.

In this section, we analyze what happens when agents form opinions about options that differ in popularity and popularity is not fixed but evolves as the agents update their opinions about the options. The central assumption of the model is that agents tend to sample options that are 'locally popular' — popular among network contacts at the time of the sampling instance. In contrast to many social learning models (Banerjee, 1992; Bikhchandani et al., 1992), this model assumes that popularity does not affect inference about the quality of the option, but only impacts sampling. There are many situations where people might select popular options even if they do like them. Studies of normative social influence and conformity have shown that people who deviate from the behaviors of others tend to be less liked and sometimes ostracized. Moreover, people might select the popular alternative because "it is better for reputation to fail conventionally than to succeed unconventionally" (Keynes 1936, p. 158). Finally, the ranking algorithms that control which options show up at the top of search results, which links are shown on our Facebook newsfeed or our Twitter feed, or which movies are shown on the top of the screen of our favorite streaming platform frequently rely on evaluations by our network contacts or people who are similar to us. Such ranking algorithms make options that are liked among our network contact more available and thus more frequently selected (Germano, Gómez, & Le Mens, 2019).

Model

We analyze the sampling behavior of *N* agents in a connected network who update their evaluations of *K* options over a set of *T* periods. In each period, an agent *i* that is randomly drawn from the population samples one of the *K* options, observes its payoff and updates their valuation of that option. The popularity of the options among *i*'s network contacts at the beginning of period *t* affects *i*'s sampling behavior in that period.

Payoff distributions of the options. There are *K* options with stable payoff distributions. We denote by f_k the density of the payoff distribution of option *k* and by u_k its mean.

Initial valuations. The valuation of option *k* at the beginning of period *t*, for agent *i*, is denoted by $V_{k,t}^i$. For all agents, the initial valuation of option *k*, $V_{k,t}^i$. consists in one random draw from its payoff distribution.

Sampling Rule. In each period, an agent *i* is randomly selected in the population. Let $P_{k,t}^i$ denote the likelihood that the agent samples option *k* in period *t*. We implement the assumption that agent *i* is more likely to sample a popular option by assuming that $P_{k,t}^i$ depends on the mean evaluation of option *k* by *i*'s neighbors–the agents to which *i* is connected (Woiczyk & Le Mens, 2021). We denote this quantity by $\bar{V}_{k,t}^i$. The agent does not fully rely on the opinions of their neighbors, however. Their own

valuations also affect the sampling probability. Accordingly, the likelihood that the agent samples option k is a logistic function of their valuations and the mean valuations of their neighbors:

$$P_{k,t}^{i} = \frac{e^{s_1 V_{k,t}^{i} + s_2 \overline{V}_{k,t}^{i}}}{\sum_{l=1}^{K} e^{s_1 V_{l,t}^{i} + s_2 \overline{V}_{l,t}^{i}}}, \quad (1)$$

where $s_1 > 0$ and $s_2 > 0$ are parameters that characterize the sensitivity of the choice probability to the valuations of the options and its popularity respectively.

Valuation Updating. We assume that the agent updates their valuation of an option based on their own experience with the option, and that this updating process is independent of the relative popularities of the options. In other words, we *'switch off' the possibility for motivated information processing* that would make the agent interpret different the same information about a popular or an unpopular alternative.

More formally, we assume that the valuation of option k at the beginning of period t, $V_{k,t}^{i}$ is equal to a weighted average of the prior valuation and the most recent payoff $x_{k,t-1}^{i}$ if it was selected in period $t-1$:

$$V_{k,t}^{i} = (1-b)V_{k,t-1}^{i} + bx_{k,t-1}^{i}, \quad (2)$$

where $b \in [0, 1]$ is the weight of the payoff and $x_{k,t-1}^{i}$ is a random draw from a binomial distribution with probability p_k. The valuations of options that were not selected do not change. Prior research has shown that the combination of a logistic choice rule and the delta rule for estimate updating provides a good fit to experimental data on sequential choice under uncertainty (Denrell, 2005).

Illustration in a setting with two agents and two options

To illustrate how the model works, we consider the case of 2 agents ('A' and 'B') learning about two uncertain options (e.g., an 'independent Catalonia' or a 'Catalonia as a part of Spain') with positive variance and unknown means u_1 and u_2. See Figure 1a. The probability agent A selects option 1 in period t is:

$$P_{1,t}^{A} = \frac{e^{s_1 V_{1,t}^{A} + s_2 V_{1,t}^{B}}}{e^{s_1 V_{1,t}^{A} + V_{1,t}^{B}} + e^{s_1 V_{2,t}^{A} + s_2 V_{2,t}^{B}}} = \frac{1}{1 + e^{-s_1 \Delta V_t^{A} - s_2 \Delta V_t^{B}}}, \quad (3)$$

where $\Delta V_t^{A} = V_{1,t}^{A} - V_{2,t}^{A}$ is A's 'opinion.' We define B's opinion similarly. Agent A is more likely to sample option 1 when their opinion favors option 1. A is also more likely to sample option 1 when B's opinion favors that option. An important feature of this choice rule is that the opinions of the two agents are *compensatory* in the sense that even if A's opinion is against option 1, A might nevertheless be likely

to sample that option if B's opinion strongly favors option 1. To see this, note that the probability that A samples option 1 in period *t* is higher than .5 whenever $\Delta V_{1,t}^{B} > -\frac{s_1}{s_2}\Delta V_{1,t}^{A}$

In the special case where B's opinion does not affect A's sampling probability, ($s_2 = 0$), this model becomes the basic 'Hot-Stove model' analyzed in the 'Hot-Stove Effect' chapter in this book. In this case, the learning process of agent A is characterized by a systematic asymmetry in error correction: Errors of underestimation are less likely to be corrected than errors of overestimation. More specifically, suppose A underestimates the value of option 1 ($V_{1,t}^{A} < u_1$). A becomes unlikely to sample option 1. By avoiding option 1, A is unlikely to obtain additional information that could help them correct this error of underestimation. Compare this to the case where A overestimates the value of option 1 ($V_{1,t}^{A} > u_1$). A is likely to sample option 1 again. By doing so, A obtains additional information about option 1 that can lead to a correction of this error of overestimation. This asymmetry in error correction implies that in every period after the first period, A has a probability lower than 50% of sampling option 1. Moreover, the expected valuation of option 1 is lower than the mean of option 1: $E[V_{1,t}^{A}] < u_1$. See the 'Hot-Stove Effect' chapter for details.

Now consider the case where A's sampling probability depends not only on A's opinion but also on B's opinion ($s_2 > 0$). Denrell and Le Mens (2017) have shown that, in this case, the valuations of option 1 by agents A and B become positively correlated. The same happens for option 2. The reason for this emergent correlation is that the compensatory nature of the choice rule influences the *joint* pattern of errors of underestimation and overestimation implied by the hot-stove effect: suppose agent A underestimates the value of option 1. A might correct this error only if A samples option 1 again. When does this happen? This happens when agent B values option 1 positively. Hence, upward corrections of underestimation errors by agent A tend to happen when agent B has a positive valuation of that option. Suppose that agent B also underestimates the value of option 1. In this case, both agents tend to sample option 2 and the joint underestimation of option 1 will persist. All-in-all, this dynamic implies that the valuations of the two options become correlated, as do the opinions ($\Delta V_{1,t}^{A}$ and $\Delta V_{1,t}^{B}$).

As an illustration, suppose that the payoff distributions of the two options are Normal with mean 0 and variance 1, that $s_1 = s_2$ and that $b = .5$. Simulations show that the correlation between the valuations of option 1 by agents A and B is initially 0 and increases over time. After 200 periods it is close to .34.[5] When $s_1 = s_2 = s$, Proposition 3 in Denrell and Le Mens (2017, p. 537) provides an explicit formula for the asymptotic correlation (the correlation after a very large number of periods):

[5] Results are based on 10,000 simulations of the model.

$$corr(V_k^A, V_k^B) = \frac{1}{1+4\frac{(2-b)}{s^2\sigma^2 b}} > 0. \quad (4)$$

This formula shows that the correlation is larger when the weight of the more recent observation (b) is larger and option choice depends more strongly on the valuations (s is larger). Moreover, the correlation is equal to 0 if the payoffs are certain ($\sigma = 0$) and larger if the payoffs are more variable (σ is large). In this case, each observation is a noisy signal of the mean of the payoff distribution of the sampled option, and estimation errors can be high.

What does this imply for the opinions and the preferences of the two agents? We will say that when the opinion of agent A favors option 1, $\Delta V_t^A > 0$, agent A 'prefers' option 1. Simulations show that both opinions and preferences become more similar over time. Initially, the opinions of the two agents are uncorrelated, and the probability of consensus (that they have the same preference) is .5. After 200 periods, the correlation between the opinions is .5 and the consensus probability is .82.

It is important to note that this model implies an emergent homogeneity in opinions and a convergence in preferences only if (1) the probability that agent i samples an option depends both on their valuations *and* the valuations of the other agent. If the choice probability depends only on the focal agent valuation ($s_2 = 0$), the two agents are subject to the hot-stove effect, but there is no interaction between them and the opinions remain uncorrelated. If the choice probability depends only on the other agent valuation ($s_1 = 0$), homogenization will not happen either. This is because, in this case, the influence of the other agent on the focal agent's opportunities for error corrections is independent of the focal agent's valuations. In this case as well, the opinions remain uncorrelated.

The 'Maximal Argument Strength' Model

This model is inspired by early work on how argument exchanges can lead to opinion extremization and polarization. Persuasive Argument Theory (PAT, Burnstein & Vinokur, 1977) proposes that people are most convinced by arguments they find unique and novel and that during deliberation more unique arguments for the position that is more popular in the group are generated. These two phenomena could explain why argument exchange leads to a tendency for the preferences of the members of the group to converge toward one position. The model we describe in this section provides a formal rendition of this intuition that we then use to examine how this mechanism could lead to polarization in a structured network. (See Mäs and Flache (2013) for a similar model.)

Model

We analyze the dynamics of valuations in a population of *N* agents connected in a network. The agents update their evaluations of two options based on the arguments they receive from others. In each period, one agent *i* is randomly drawn. This agent samples arguments in support of the two options from their network neighbors. When prompted, a neighbor provides an argument in support of the option they prefer. After collecting the arguments, agent *i* updates their valuations of the two options based on the strongest arguments they heard in support of each option.

Initial valuations. For all agents, the initial valuation of option *k* consists in one random draw from a uniform distribution: $V_{k,1}^{i} \sim U(0,1)$.

Argument Generation. Suppose agent *i* samples arguments supporting each option from their network neighbors. We implement the assumption that agents try to persuade others of their positions by assuming they only generate arguments in support of their preferred option. Consider option *k*. The agents who generate arguments for option *k* are all the agents that are the neighbors of agent *i* for which option *k* is the preferred option. Therefore, if $\eta_{k,t}^{i}$ is the set of *i*'s neighbors who prefer option *k*, agent *i* will sample $|\eta_{k,t}^{i}|$ arguments in favor of this option from their neighbor. We also assume that *i* generates one argument of their own.[6] Each argument is a random draw from the uniform $U(0,1)$ distribution. This captures the possibility that although the agent's evaluation of an option is very high, the argument they provide can be perceived as weak by agent *i*.

Consistent with Persuasive Argument Theory, we assume that *i* is most influenced by the most 'persuasive' argument in support of an option, defined as the argument with the largest value on the [0, 1] interval. More formally, the strength of the persuasive argument for option *k* is:

$$\mathcal{A}_{k,t}^{i} = \max_{j \in \eta_{k,t}^{i}} a_{k,t}^{j}, \tag{5}$$

where $a_{k,t}^{j} \sim U(0,1)$.

Valuation updating. The valuation updating rule differs from the one used in the 'Asymmetric Hot-Stove' model. Here, the agent does not make choices that affect the information they sample. Instead, the agent updates their valuations of all available options based on the strength of the most persuasive argument sampled for each option, using the delta rule:

$$V_{k,t}^{i} = (1-b)V_{k,t-1}^{i} + b\mathcal{A}_{k,t}^{i}, \tag{6}$$

[6] This is mostly motivated by technical considerations regarding the simulation of what happens when alternative *k* is preferred by none of the agents.

where $b \in [0, 1]$ is the weight of the new argument.

Illustration in a setting with two agents and two options

Consider again a simple network with just two agents ('A' and 'B'). We first consider the case where, initially, the two agents both prefer option 1, without loss of generality ($\Delta V_t^A > 0$ and $\Delta V_t^B > 0$).

Suppose agent A is the focal agent in period 1. A samples two arguments about their preferred option (option 1): the argument they generate and the argument B generates. By contrast, A samples just one argument about their least preferred option (option 2). Because the strength of the relevant argument depends on sample size (the maximum of two independent realizations of a random variable tends to be higher than one realization), it is likely that the persuasive argument for option 1 is stronger than the persuasive argument for option 2: $\mathcal{A}_{A,1}^1 > \mathcal{A}_{A,1}^2$. There is a 2/3 probability that this is the case. Because option valuation at beginning of period 2 is the weighted average of valuation at the beginning of the period and of the strength of the persuasive argument, there is at least a 2/3 probability that, at the beginning of period 2, agent A prefers option 1. Suppose that in period 2, the focal agent is B. Since A prefers option 1, agent B will sample two arguments in favor of option 1 and one argument in favor of option 2. By the same logic as that applied to A, B is likely to keep preferring option 1.

Now, consider the case where the two agents have different initial preferences. Without loss of generality, we assume that A prefers option 1 and B prefers option 2. Suppose that A is the focal agent in period 1. B's preference implies that B will provide an argument for option 2 that will be considered together with the one argument for each option generated by A. By the logic outlined above, this will lead to a likely increase in the evaluation of option 2 for A. There are two possible scenarios.

In the first scenario, the increase leads to a change in A's preferences in favor of option 2. This leaves us in the situation described in the previous paragraph. In the second scenario, A's preference does not change, yet with a 2/3 probability, A's opinion will become less unfavorable to option 2 than before. In the next period, B samples arguments. B's opinion will likely shift (to some extent at least) toward option 1. Then the question the same question can be asked about B: did this update lead to a preference reversal in favor of option 1? Eventually, the preference of one of the agents will flip and both agents will prefer the same option. The reinforcing dynamics described above will lead to a stochastically stable agreement among the agents.

Just as the Asymmetric Hot-Stove model, this model does not lead to a deterministic 'lock-in.' In each period, the agents update their valuations of the options based on arguments that could lead to a preference reversal. When the weight of new arguments (b) is low, convergence becomes quite stable. Simulations of the model with $b = .05$ show that, after 200 periods, the consensus probability is .995 and

the correlation between the agents' opinions is .88. By contrast, if the weight of new arguments is high, opinion homogenization is milder and preference convergence less likely. With $b = .5$, after 200 periods, the consensus probability is .75 the correlation between the agents' opinions is .49. It is noteworthy, however, that preference convergence tends to be much stronger than with the 'Asymmetric' hot-stove model.

The 'Feedback Sampling' Model

This model relies on two central assumptions. First, when deciding among options, agents seek positive feedback (Thorndike, 1927). In the social media setting, this assumption is consistent with recent evidence, collected by Facebook and LinkedIn, about the reasons users post content on social media (Eckles et al., 2016; Chen et al., 2017). For example, researchers from LinkedIn experimented with manipulating the news feed display order to increase the visibility of content shared by users who were not yet decided if they wanted to continue posting on the platform (Chen et al., 2017). The users whose posts were promoted in their friends' news feeds obtained more feedback and remained more engaged on the platform. Second, agents give positive feedback to choices they also like. They give positive feedback to messages on issues they care about, or to opinion statements aligned with their own opinions. This assumption is consistent with existing evidence that users of social media give more 'likes' to like-minded content (Garz, Sood, Stone, & Wallace, 2018).

Model

We analyze the dynamics of opinions in a population of N agents connected in a network who make a series of choices between a fixed set of K options and update their valuations of the options based on the feedback they receive from the agents to which they are connected (their network 'neighbors'). In each period, one agent i is randomly drawn. This agent selects an option and receives feedback from one other agent j, randomly drawn among their network neighbors.

Initial valuations. For all agents, the initial valuation of option k consists in one random draw from a uniform distribution: $V_{k,1}^{i} \sim U(0,1)$.

Valuation updating. The valuation updating rule is the same as in the Asymmetric Hot-Stove model (eq, 2) except for the fact that it is updated based on the feedback $F_{k,t-1}^{i} \in \{0,1\}$. The valuations of options that were not selected do not change.

Choice Rule. At each period, an agent i is randomly drawn in the population. Let $P_{k,t}^{i}$ denote the likelihood that the agent selects option k in period t. We implement the assumption that agents seek

positive feedback by assuming that $P_{k,t}^i$ increases with $V_{k,t}^i$. More precisely, we assume a logistic choice rule:

$$P_{k,t}^i = \frac{e^{sV_{k,t}^i}}{\sum_{l=1}^{K} e^{sV_{l,t}^i}}, \tag{7}$$

where $s > 0$ is a parameter that characterizes the sensitivity of the choice probability to the valuations of the options. When *s* is large, the agent almost always selects the option with the highest valuation. When *s* is close to 0, choice is almost random.

Feedback. The feedback giver, *j*, is a random network neighbor of agent *i* — an agent connected to *i*. It is important to note that agents who are not connected to *i* do not provide any feedback.

The feedback giver is more likely to provide a 'like' when they value highly the option chosen by agent *i*. We implement this assumption by assuming that the probability of positive feedback is a logistic function of the option valuations by the feedback giver, *j*:

$$\phi_{k,t}^j = \frac{e^{\lambda V_{k,t}^j}}{\sum_{l=1}^{K} e^{\lambda V_{l,t}^j}}, \tag{8}$$

where *k* is the option chosen by agent *i* and $\lambda > 0$ is a parameter that characterizes the sensitivity of the feedback probability to the valuations of the alternatives. When λ is large, the feedback giver is very likely to give a 'like' when they like the chosen option and very unlikely to give a like when they do not like it. When λ is close to 0, feedback is almost random.

Illustration in a setting with two agents and two options

To understand how the model can give rise to opinion homogeneity, we first consider the case of 2 agents ('A' and 'B') learning about two uncertain options. For simplicity, we assume an extreme choice rule (the agent always selects their preferred option, $s = \infty$) and an extreme feedback rule (the feedback giver *j* gives a 'like' to the focal agent *i* if *i* selected the option *j* values the most, and does not give a 'like' otherwise, $\lambda = \infty$). Note that the valuations are between 0 and 1. This is because the initial valuation is assumed to be a random draw in this interval and the valuation updating rule (equation 2)

implies that the valuation of an option gets closer to 0 or 1 as a function of the observed feedback (each feedback instance is equal to 0 or 1).

Suppose that, initially, there exists a consensus such that both agents prefer option 1. Suppose, moreover, that A is drawn to make a choice in period 1. The preferences of A and B imply that A selects option 1, and B gives A a 'like.' Agent A's valuation for option 1 thus increases and becomes closer to 1. Suppose, without loss of generality, that B makes the period 2 choice. B prefers option 1 and thus selects option 1. Because A prefers option 1, A gives B a 'like.' Hence, B's valuation for option 1 increases. After the first two periods, both agents still prefer option 1. A recursive argument implies that this pattern persists in every period: both agents prefer option 1 in every period. A similar dynamic occurs if both agents initially prefer option 2. In this case, both agents prefer option 2 in every period. In summary, an initial consensus is persistent in every period.

What happens if A initially prefers option 1 whereas B initially prefers option 2? A selects option 1 and does not get a like (because B prefers option 2). Agent A's valuation of option 1 thus goes down. If A's valuation of option 1 becomes so low that it leads to a preference reversal in favor of option 2, both agents prefer option 2 at the beginning of period 2. There is consensus, and the reasoning of the previous paragraph implies that the consensus persists in all subsequent periods. Initially, there was no consensus, but a consensus emerged in period 2 and remains in all subsequent periods.

Consider now the case where A's preference does not shift in period 1. Suppose, without loss of generality, that B makes the choice in period 2. B prefers option 2 and thus selects that option. Because A prefers option 1, A does not give a like to B. If B's valuation of option 2 becomes lower than their valuation of option 1, B's preference changes in favor of option 1. Both agents prefer option 1 at the beginning of period 2. There is a consensus, and the above reasoning for the case with an initial consensus implies that the consensus is persistent. If B's preferences do not change, the situation at the beginning of period 3 is similar to what it was at the beginning of period 1.

When the two agents start without consensus, the only uncertain aspect in the dynamics of valuations and preferences is the period when a consensus first happens. With probability 1, it will happen at some point, but it is not possible to predict ex-ante on which option the agents will converge.

Consistent with the dynamics of preferences, the opinions of the two agents become correlated. They are initially independent, but after 200 periods, the correlation between the opinions is close to .90. When the choice rule and the feedback rules are probabilistic, (s and λ are finite) similar dynamics of opinion homogenization and preference convergence unfolds. Even though the persistence of a consensus is no longer deterministic, simulations show that as soon as feedback givers are somewhat discriminant in the way they give feedback (λ is not close to 0), then the consensus probability after 200 periods is high (e.g., with $b = .5$, $s = 5$, this is equal to .91 with $\lambda = 2$ and to .99 with $\lambda = 3$). When choice is random (s

= 0), homogenization and convergence are a bit weaker than in the previous case but remain very strong (e.g., with $b = .5$, $s = 0$, the consensus probability is equal to .85 with $\lambda = 2$ and to .96 with $\lambda = 3$). Finally, opinion homogenization and preference convergence remain high even if the belief updating weight is low (e.g., with $b = .05$, $s = 5$, the consensus probability is .93 with $\lambda = 2$ and to .98 with $\lambda = 3$).

Local Opinion Homogenization and Between Group Polarization

To explore the possibility of homogenization of opinions in densely connected social networks and between-group polarization, we analyze a two-group network of 10 agents that belong to groups of 5 agents each ({A,B,C,D,E} and {F,G,H,I,J}, see Figures 1b-d). We define a group as a set of agents that are densely connected to each other but weakly connected to the rest of the network. In our simulations, we assume that the 5 agents of each group are connected to all other members of the group. We vary the density of connections between the groups by considering the cases with 0 links, 1 link, and 5 links. We simulated the dynamics of valuations over 1,000 periods (about 100 choices by each agent). Numerical estimates are based on 10,000 simulations of the models with the same baseline parameters as in the previous section. For all three models, the valuations become correlated and preferences converge within-group.

When the two groups are disconnected, the within-group dynamics are independent from each other. This implies that whenever consensus emerges in the two groups, it happens on the same option (global consensus) about 50% of the time and toward different options (complete polarization) about 50%. An important insight resulting from these simulations is that the process of within-group convergence is *sufficient* to create an overall tendency for the groups to become more distinct in terms of preferences – preferences become more polarized. This is because we did not assume there was some 'repulsive' forces between agents of the two groups (we do so in the next section).

To quantify polarization, we denote by $\mathcal{P}_A$ ($\mathcal{P}_B$) the proportion of agents who prefer option 1 in group A (group B). We are interested in the between-group gap in preferences $\Delta\mathcal{P} = |\mathcal{P}_A - \mathcal{P}_B|$. This gap is initially equal, on average, to .25. It grows over time to become close to .5 with all three models (it is a bit lower at .45 with the Asymmetric Hot-Stove model, which is not surprising given within-group consensus is less likely to emerge with this model). Overall, the process of within-group homogenization thus leads to a general tendency for the groups to become more dissimilar.

When the preferences of the members of the two groups start to converge on different options, it is unlikely that global consensus will be achieved. This is because social influence operates via network contacts who belong to the same group – it is local. A useful metric is the 'average local support.' For each

agent, we compute the proportion of their network contacts that have the same preference as them. We call this the 'local support' of the preference of the focal agent. We then average the local support across all agents in the network to obtain the 'average local support'. Because initial valuations are random and independent for each agent, the average local support is initially equal to .50. Simulations reveal that it increases quickly. After 1,000 periods, the mean average local support is higher than .75 for all models and essentially 1 for the Feedback Sampling and Maximal Argument Strength models. This means that agents are surrounded by other agents with similar preferences to them. This is the case even when there is no within-group consensus. The fact that local support is so high implies that if the two groups tend to favor distinct options at some point, it is extremely unlikely that global consensus will ever happen.

With some between-group links, the connections make it more likely (than without any connection) that, early on in the process, both groups happen to have a majority of agents who prefer the same option. When this happens, the reinforcing dynamics that mostly operate within the group imply that both groups will tend to converge toward the same option. In other words, the probability of global consensus increases, and the tendency toward polarization decreases as compared to the 'disconnected' group case.

The initial 'coupling' of the within-group dynamics is stronger when there are more between-group connections. Accordingly, for all three models, the probability of global consensus (all agents preferring the same option) is higher with one between-group link than with no link, and it is still higher with five between-group links. The consensus probabilities for the three models and 0/1/5 links are .11/.15/.2 (Asymmetric Hot-Stove), .49/.54/.74 (Maximal Argument Strength), .50/.58/.92 (Feedback Sampling). Similarly, the between preference gap becomes lower with more links. The gaps for the three models and 0/1/5 links are .46/.41/.27 (Asymmetric Hot-Stove), .50/.46/.26 (Maximal Argument Strength), .50/.40/.19 (Feedback Sampling).

Finally, it is worth noting that even when within-group consensus is not achieved, some level of local convergence does occur. To characterize this, it is useful to compute the average local support over the simulation runs for which global consensus did not emerge. Even in these cases, the average local support is higher than .67 with all three sampling models and 0 to 5 between-group links. This means that agents are connected to twice as many other agents with the same preference as them as compared to agents with different preferences. In other words, agents become surrounded by 'like-minded' others.

Local Opinion Homogenization and Polarization in a Network of Two Groups with Distinct Identities

The assumptions of our sampling models regarding the nature of social interactions between network neighbors lead to opinion homogenization among densely connected agents. Interactions do not have to contribute to homogenization, however. Conflicts are an important part of human social

interaction and can result in aggression, especially when group identity is involved (Densley & Peterson, 2018). Models of 'repulsive' social influence have recognized this possibility by allowing encounters with someone an agent disagrees with to result in even larger opinion differences (Baldassarri & Bearman, 2007; Flache & Macy, 2011). The main psychological process proposed as the source of repulsive influence is the desire of people to differentiate themselves from dissimilar or disliked others (Brewer, 1991). Others can be disliked because they have different options from the agents (Rosenbaum, 1986) or because they belong to a group with a different identity — an 'out-group' (Tajfel, Billig, Bundy, & Flament, 1971). In this section, we incorporate these ideas into each of the three sampling models. We assume there are two groups in the network (see Fig. 1e). We denote the group to which the focal agent *i* belongs as 'the in-group' and the group *i* does not belong to as 'the out-group.' Agents interact differently with members of the 'in-group' than with members of the 'out-group.'

Asymmetric Hot-Stove Model. The choice of option by the focal agent is influenced differently by members of the two groups. Agent *i* tends to select options that are popular in the in-group but that are also unpopular in the out-group. In other words, the agent tries to avoid options that are popular in the out-group. This could happen when agents see their actions as identity signals, and want to preserve a distinct identity (e.g., Brewer, 1991). We implement these assumptions by assuming that $P^i_{k,t}$ depends on the mean evaluation of option *k* by *i*'s neighbors who are in the in-group $\overline{V}^i_{k,t,in}$ and on the mean evaluation of by *i*'s neighbors who are in the out-group ($\overline{V}^i_{k,t,out}$) as follows:

$$P^i_{k,t} = \frac{e^{sV^i_{k,t}+s_{in}\overline{V}^i_{k,t,in}-s_{out}\overline{V}^i_{k,t,out}}}{\sum_{l=1}^{K} e^{sV^i_{l,t}+s_{in}\overline{V}^i_{l,t,in}-s_{out}\overline{V}^i_{l,t,out}}}, \quad (9)$$

where $s > 0$, $s_{in} > 0$, and $s_{out} > 0$ are parameters that characterize the sensitivity of the choice probability to the valuations of the options and its popularity among in-group and out-group neighbors respectively. In the simulations reported below, we take $s = s_{in} = 5$, $s_{out} = 5$ and $b = .5$

Maximal Argument Strength Model. Here we consider the cases where there is some 'mistrust' between the members of the groups, maybe because they have a competitive relationship. Consider an agent *i* who evaluates an option *k* based on arguments produced by in-group and out-group members. Let $\mathcal{A}^i_{k,t,in}$ denote the most persuasive argument from the in-group and $\mathcal{A}^i_{k,t,out}$ be the most distinctive argument produce by out-group members about option *k*. We assume that arguments produced by the out-group influence the agent negatively because the focal agent suspects that the out-group member is trying

to mislead them by engaging in strategic behavior. This assumption is also consistent with research that has documented a backfiring effect of exposure to groups of distinct identities (Nyhan & Reifler, 2015; Taber & Lodge, 2006). The valuation of the available options is updated as follows:

$$V_{k,t}^{i} = (1 - b_{in} + b_{out})V_{k,t-1}^{i} + b_{in}\mathcal{A}_{k,t,in}^{i} - b_{out}\mathcal{A}_{k,t,out}^{i}, \quad (10)$$

where $b_{in} \in [0,1]$ is the updating weight for arguments produced by in-group members and $b_{out} \in [0,1]$ is the updating weight for arguments produced by out-group members. In the simulations reported below, we take $s = 5$, $b_{in} = .05$ and that $b_{out} = .05$

Feedback Sampling Model. This model can also be adapted to reflect situations of strategic behavior. There are two possible approaches to capture this kind of setting: focusing on how the feedback recipient interprets feedback, or focusing on how the feedback givers provide feedback. In the former case, we can adopt an approach similar to that used to adapt the Maximal Argument Strength Model. Because the agent wants to be distinct from the out-group and interprets feedback as a trustful signal of appraisal, they will decrease their evaluations of options that are endorsed by the out-group—they give negative weight to feedback by members of the out-group.[7] In the latter case, we assume that feedback givers try to influence members of the other group to select options different from them. If they believe the feedback recipients will interpret their feedback at 'face-value,'[8] they will simply give a 'like' when the other agents select their least preferred alternative. To implement this in our model, we assume that the feedback rule is the same as before (eq. 8) when the feedback giver is from the same community as the focal agent and that feedback is 'flipped' when they are from a different community: they give a 'like' whenever they would not have given a 'like' to an agent from their community, and they fail to give a 'like' whenever they would have given a 'like.' In the simulations reported below, we implement the second approach and take $b = .05$, $s = 5$ and $\lambda = 5$.

Results. We simulated the sampling models for 1,000 periods in a setting with 5 between-group links. With all three models, the opinions of agents in the same group become positively correlated (similar to what was obtained without distinct identities), whereas the opinions of agents of the two groups now become negatively correlated (they were uncorrelated or positively correlated in the previous simulation set). What changes in comparison to the setting without identities is that the probability of a global consensus becomes much lower (Asymmetric Hot-Stove: .19 → 0; Maximal Argument Strength: .74 → .04; Feedback Sampling: .75 → .14). Relatedly, the gap in preferences between the two groups ($\Delta\mathcal{P}$) is

[7] It would also be possible to consider the case where feedback recipients suspect the out-group members to give feedback that does not reflect their true preferences, but this opens the door to complications associated to multi-level reasoning in games

[8] This is most realistic in settings where feedback recipients do not easily know the identity of the feedback givers, for example with 'likes' on social media posts

much higher than in the setting without identity and becomes high in all cases (Asymmetric Hot-Stove: .27 → .57; Maximal Argument Strength: .26 → .82; Feedback Sampling: .19 → .81).

The most interesting finding from these simulations is obtained with versions of the model where only negative influence occurs (Asymmetric Hot-Stove: $s_{in} = 0$; Maximal Argument Strength: $b_{in} = 0$; Feedback sampling: no feedback for to agents in the same group). In this case, the opinions of agents that belong to different groups become negatively correlated, but there is no clear tendency for polarization (the probability that the majorities of the two groups prefer different options is lower than .5 for all three models). This demonstrates that within-group homogenization is a necessary component for polarization to emerge with these sampling models.

Implications

Consider again Florian, the German exchange student in Barcelona. As we have shown throughout the chapter, the fact that his classmates are pro-independence makes him less likely to sample anti-independence information, more likely to sample more persuasive arguments about the pro-independence position and to be rewarded by his classmates when he expresses a pro-independence stance. All this will likely lead Florian to speak in favor of an independent Catalonia to his friends back in Germany, believing he is sharing the dominant view.

Our findings about the strength of 'local support' produced by the three models show that it is possible that Florian's opinion be not as widely shared as he might think. Florian's opinion could be largely unpopular in the general population even though it is popular among Florian's direct contacts. This inconsistency between local and global support is more likely to emerge if the group to which Florian belongs is somewhat disconnected from the rest of the population (Galesic, Olsson, & Rieskamp, 2012, 2018). When this happens, Florian's opinion might even become extreme. Our numerical illustrations relied on simplistic model of opinion extremization in which there were just two choice alternatives. It is easy, however, to extend the models to a setting with more alternatives. Such extension would provide a sampling explanation for how an extreme opinion can emerge in an isolated group in a network and remain stable over time. Such groups could consist in online forums and private groups that, as Sunstein argues (Sunstein, 2018), become sources of extreme views and conspiracy theories. A timely example consists of the anti-vaccine movement, which seems to have had negative effects on the COVID-19 vaccination campaigns in several countries.

In addition to suggesting that simple mechanisms of information sampling could contribute the emergence of polarized and extreme opinions, the sampling approach discussed in this chapter offers a different perspective on how society can limit polarization. Existing theories of polarization are largely based on mechanisms that invoke motivated information processing. These imply that, to correct the

tendency for local opinion homogeneity, one needs to correct how information is processed by the individual. The sampling approach, by contrast, implies that the correction should focus on the nature of the information sampled by agents, before any processing by their minds happens. As our simulations show, the connectivity of the network has a crucial influence on the level of opinion polarization. The sampling approach, thus, suggests a seemingly simple solution: increase the diversity of information to which people are exposed. The models discussed in this chapter offer specific suggestions regarding how sampling diversity can be increased. Because the Asymmetric Hot-Stove and Maximal Argument Strength models are concerned with sampling of information through newsfeeds, they suggest that opinion polarization can be reduced by simply expanding the user's newsfeed and set of connections (which influences the information sent onto the newsfeed) to be more diverse. The Feedback Sampling model suggests that a possible way to combat polarization is to reduce feedback's visibility and promote discussion in the comments where users can engage in a more argument-based conversation.

Even though they are ostensibly simple, these potential solutions would be difficult to implement in practice. First, most social media platforms provide limited control to their users over the information that reaches them via their newsfeed and over the nature of the feedback provided to them. Moreover, the newsfeed algorithms are often proprietary. Second, more diverse information sampling might go against an individual's hedonic goals: the tendency to seek positive experiences. Being challenged about one's opinions or receiving negative feedback are rarely positive experiences. The threat of negative experience could thus discourage individuals from seeking to broaden their information sampling, and they might stop using the social media platform altogether. This, in turn, encourages social media companies to design information environments that maximize the hedonic quality of user experiences. This results in curated newsfeeds that are skewed towards the arguments popular among network contacts and asymmetric feedback structures where expressing support is much easier than disagreement. Facebook has experimented with the approach that consists in making different perspectives on an issue easier to sample by creating 'Related articles'.[9] Yet, making information easy to sample does not necessarily imply that people will sample it, especially if their identity affects their sampling behavior. As shown by our simulations of models with group visible identities, identity exacerbates the effects of information sampling and increases probability of polarization. More generally, the resolution of the tension between the frequently hedonic goals of individuals and society's need for informed citizens remains a formidable challenge for public policy.

Even though we mostly discussed the effects of information sampling on the public, the models in this chapter have implications for understanding the behavior of politicians. Imagine a politician who is

[9] https://about.fb.com/de/news/2017/09/update-zu-den-wahlen/

considering new legislation. The sampling models in this chapter propose that the composition of politicians' social circle affects the information that politician samples to inform their opinion about the policy. Research shows that politicians are more likely to interact with fellow party members and supporters that are on the same side of the political spectrum (Barberá et al., 2019). Then, according to the 'Feedback sampling model', a politician would express support for the policy when it is received well by their social contacts. This is exactly what Schöll, Gallego, and Le Mens (2021) found by analyzing the behavior of Spanish politicians on Twitter: they tended to post more about topics that previously received a lot of likes and retweets. In addition to social media feedback, other mechanisms can affect politicians' opinions. A politician will be relatively more exposed to information popular among their social circle and will find it more persuasive. This skew towards information in support of a specific opinion can result in the politician believing that the policy has stronger public support than it actually has. Therefore, these sampling models contribute to explaining why politicians misperceive public opinion (Broockman & Skovron, 2018), are more sensitive to the influence of the wealthy (Gilens & Page, 2014), and sometime support policies that contradict public consensus.

References

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

Banerjee, A. V. (1992, 8). A Simple Model of Herd Behavior. *The Quarterly Journal of Economics*,*107*(3), 797–817. doi:10.2307/2118364

Barberá, P., Casas, A., Nagler, J., Egan, P. J., Bonneau, R., Jost, J. T., & Tucker, J. A. (2019, 11). WhoLeads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass

Public Using Social Media Data. *American Political Science Review*, *113*(4), 883–901. doi: 10.1017/S0003055419000352

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992, 10). A theory of fads, fashion, custom, and culturalchange as informational cascades. *Journal of Political Economy*, *100*(5), 992–1026. doi: 10.1086/261849

Brewer, M. B. (1991, 10). The Social Self: On Being the Same and Different at the Same Time.*Personality and Social Psychology Bulletin*, *17*(5), 475–482. doi: 10.1177/0146167291175001

Broockman, D. E., & Skovron, C. (2018, 8). Bias in perceptions of public opinion among political elites.*American Political Science Review*, *112*(3), 542–563. doi: 10.1017/S0003055418000011

Burnstein, E., & Vinokur, A. (1977). Persuasive argumentation and social comparison as determinants ofattitude polarization. *Journal of Experimental Social Psychology*, *13*(4), 315–332. doi: 10.1016/0022-1031(77)90002-6

Burt, R. S. (2009). *Structural holes: The social structure of competition*. Harvard university press. Chen, G., Chen, B.-C., & Agarwal, D. (2017). Social incentive optimization in online social networks. In *Proceedings of the tenth acm international conference on web search and data mining* (pp. 547–556).

DeGroot, M. H. (1974). Reaching a Consensus. *Journal of the American Statistical Association*, *69*(345),118–121. doi: 10.1080/01621459.1974.10480137

Denrell, J. (2005). Why most people disapprove of Me: Experience sampling in impression formation.*Psychological Review*, *112*(4). doi: 10.1037/0033-295X.112.4.951

Denrell, J., & Le Mens, G. (2007). Interdependent sampling and social influence. *Psychological Review*,*114*(2), 398–422. doi: 10.1037/0033-295X.114.2.398

Denrell, J., & Le Mens, G. (2017, 2). Information Sampling, Belief Synchronization, and CollectiveIllusions. *Management Science*, *63*(2), 528–547. doi:10.1287/mnsc.2015.2354

Densley, J., & Peterson, J. (2018, 2). *Groups Aggression* (Vol. 19). Elsevier B.V. doi: 10.1016/j.copsyc.2017.03.031

Eckles, D., Kizilcec, R. F., & Bakshy, E. (2016). Estimating peer effects in networks with peer encouragement designs. *Proceedings of the National Academy of Sciences*, *13*(27), 7316–7322.

Flache, A., & Macy, M. W. (2011). Small Worlds and Cultural Polarization. *The Journal of Mathematical Sociology*, *35*(1-3), 146–176. doi: 10.1080/0022250X.2010.532261

Friedkin, N. E. (1999). Choice Shift and Group Polarization. *American Sociological Review*, *64*(6), 856. doi: 10.2307/2657407

Galesic, M., Olsson, H., & Rieskamp, J. (2012, 12). Social Sampling Explains Apparent Biases in Judgments of Social Environments. *Psychological Science*, *23*(12), 1515–1523. doi: 10.1177/0956797612445313

Galesic, M., Olsson, H., & Rieskamp, J. (2018, 4). A sampling model of social judgment. *Psychological Review*, *125*(3), 363–390. doi: 10.1037/rev0000096

Garz, M., Sood, G., Stone, D. F., & Wallace, J. (2018). What Drives Demand for Media Slant? *Working Paper*.

Germano, F., Gómez, V., & Le Mens, G. (2019). The few-get-richer: A surprising consequence of popularity-based rankings. In P*roceedings of the world wide web conference, www 2019.* doi: 10.1145/3308558.3313693

Gilens, M., & Page, B. I. (2014, 9). Testing theories of American politics: Elites, interest groups, and average citizens. *Perspectives on Politics*, *12*(3), 564–581. doi: 10.1017/S1537592714001595

Isenberg, D. J. (1986). Group Polarization. A Critical Review and Meta-Analysis. *Journal of Personality and Social Psychology*, *50*(6). doi: 10.1037/0022-3514.50.6.1141

Iyengar, S., Sood, G., & Lelkes, Y. (2012). *Affect, not ideology: A social identity perspective on polarization* (Vol. 76) (No. 3). doi: 10.1093/poq/nfs038

Latané, B., Nowak, A., & Liu, J. H. (1994). Measuring emergent social phenomena: Dynamism, polarization, and clustering as order parameters of social systems. Behavioral Science, 39(1), 1–24. doi: 10.1002/bs.3830390102

Mäs, M., & Flache, A. (2013). Differentiation without Distancing. Explaining Bi-Polarization of Opinions without Negative Influence. PLoS ONE, 8(11), e74516. doi: 10.1371/journal.pone.0074516

McGarty, C., Turner, J. C., Hogg, M. A., David, B., & Wetherell, M. S. (1992). Group polarization as conformity to the prototypical group member. British Journal of Social Psychology, 31(1), 1–19. doi: 10.1111/j.2044-8309.1992.tb00952.x

Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. Journal of Personality and Social Psychology, 12(2). doi: 10.1037/h0027568

Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. Psychological Bulletin, 83(4), doi: 10.1037/0033-2909.83.4.602

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

Nyhan, B., & Reifler, J. (2015, 1). Does correcting myths about the flu vaccine work? An experimental evaluation of the effects of corrective information. Vaccine, 33(3), 459–464. doi: 10.1016/j.vaccine.2014.11.017

Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. Penguin UK.

Rosenbaum, M. E. (1986). The repulsion hypothesis: On the nondevelopment of relationships. Journal of Personality and Social Psychology, 51(6), 1156–1166.

Schöll, N., Gallego, A., & Le Mens, G. (2021). How politicians learn from citizens' feedback: The case of gender on Twitter. *Working Paper.*

Sunstein, C. R. (2018). #Republic. Princeton University Press. doi: 10.1515/9781400890521

Taber, C. S., & Lodge, M. (2006, 7). Motivated skepticism in the evaluation of political beliefs. AmericanJournal of Political Science, 50(3), 755–769. doi: 10.1111/j.1540-5907.2006.00214.x

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971, 4). Social categorization and intergroup behaviour. European Journal of Social Psychology, 1(2), 149–178. doi: 10.1002/ejsp.2420010202

Thorndike, E. L. (1927). The law of effect. The American journal ofpsychology, 39(1/4), 212–222.

Turner, J. C., Wetherell, M. S., & Hogg, M. A. (1989). Referent informational influence and group polarization. British Journal of Social Psychology, 28(2), 135–147. doi: 10.1111/j.2044-8309.1989.tb00855.x

Woiczyk, T. K. A., & Le Mens, G. (2021). Evaluating categories from experience: The simple averaging heuristic. Journal of Personality and Social Psychology, 121(4), 747–773. https://doi.org/10.1037/pspa0000231