Shared Models in Networks, Organizations, and Groups

Preliminary and Incomplete

Joshua Schwartzstein and Adi Sunderam*
Harvard Business School

August 13, 2021

Abstract

Why did the market rise yesterday? What are the implications of the latest school shooting? Why did a particular employee get promoted? To answer such questions, we often exchange models, stories, narratives, and interpretations with others. This paper provides a framework for thinking about such social exchanges of models. The key assumption, following Schwartzstein and Sunderam (2021), is that when people are exposed to multiple interpretations of the data they adopt the one that provides the best explanation. A key implication is that within a network interpretations evolve. This evolution driven by social learning *hardens* reactions to data that are open to interpretation: following the exchange of models, people are more convinced they are able to explain the data. Thus, people in different networks can not only end up with vastly different beliefs, but also in a sense be puzzled by the fact that others outside their network have different beliefs. For certain network structures, we show that social learning also *mutes* reactions to data that are open to interpretation: following the exchange of models, people's beliefs are closer to their priors than before this exchange. In addition to studying fixed networks, we also consider how firm managers, politicians, and other agents are able to shape the communication network to their advantage. Agents who benefit from muting or from people sharing the same model will encourage a robust exchange of interpretations; agents who instead want new data to change behaviors will try to limit the exchange of interpretations, especially interpretations that suggest the data are not surprising. We apply the framework to consider the goal and structure of meetings in organizations, as well as the evolution and persistence of myths in social networks.

*We thank Tristan Gagnon-Bartsch, Simone Galiperti, Robert Gibbons, Andrei Shleifer, Mario Small, and seminar participants at UC San Diego, Stanford GSB, and University of Zurich for helpful comments.

1 Introduction

We make sense of the world together. Why is the unemployment rate better than expected? Why did one employee receive a promotion while another did not? Why did a political candidate underperform her polls? Why is the price of a certain stock shooting up? In response to such questions, we share not only information but also interpretations. Unemployment numbers are better than expected "because the economy is doing better than expected" or because "there was a one-time blip in certain sectors". One candidate received a promotion over another because "she is uniquely qualified" or "the employer is sending a signal that her particular skills are generally valued by the organization". The stock price is shooting up because of "fundamentals" or "dumb money". What is the outcome of this exchange of interpretations? Do we end up with interpretations closer to the truth than where we started? How does who we talk to influence what we come to believe? And how might an interested party like a firm manager seek to influence patterns of communication to shape ultimate interpretations?

This paper presents a formal framework for thinking about such social exchanges of interpretations. The basic ingredients of the model closely follow Schwartzstein and Sunderam (2021). Everyone shares a common prior μ_0 over states of the world ω and observes a common, public history h relevant to updating their beliefs before taking an action to maximize their expected utility. Aspects of the history are open to interpretation, meaning that people are willing to entertain many different interpretations of the same data. Interpretations are represented by models, which we formalize as likelihood functions that link the history to states. In other words, interpretations capture the ways people use the history to update their beliefs. When people are exposed to multiple interpretations of the data, they adopt the one that provides the best fit to the data, fixing prior beliefs. People have a default interpretation d, represented by likelihood function $\pi_d(h|\omega)$, and come up with a single alternative interpretation—their initial reaction to the data—that they adopt if it is more compelling, i.e., it fits the data plus their prior better, than their default interpretation.

In contrast to standard social learning models where people learn from others' actions or signals (e.g., reviewed in Golub and Sadler (2016)), in our framework everyone shares the same information but learns from others' interpretations. People are exposed to the interpretations of others within their network and settle on the interpretation they are exposed to that is most compelling. Formally, person i adopts the model she is exposed to m (represented by likelihood function $\pi_m(h|\omega)$) if

$$m \in \arg \max_{\tilde{m} \in \left\{d, m_i'\right\} \cup M_i} \underbrace{\Pr(h|\tilde{m}, \mu_0)}_{= \int \pi_{\tilde{m}}(h|\omega) d\mu_0(\omega)},$$

where m_i' represents the model person i comes up with initially and M_i is the set of models the person is exposed to in her network. In Bayesian terms, the person acts as if she has a flat prior

over the models she is exposed to and then selects the model with the highest associated posterior probability. More intuitively, this assumption loosely captures ideas from the social sciences about what people find persuasive, including that people favor models which (i) have high "fidelity" to the data as emphasized in work on narratives (Fisher 1985); (ii) help with "sensemaking" as discussed in work on organizational behavior and psychology (Chater and Loewenstein (2016); Weick (1995)); and (iii) make the past feel more predictable (Schulz and Sommerville (2006); Gershman (2019)).

To see some basic implications of this formulation, consider right-leaning voters who are trying to assess the outcome of an election—both in terms of who won (i.e,. received the highest certified vote tally) and whether the election was fair. (The analysis of left-leaning voters is symmetric.) Voters' priors are that the left-leaning and right-leaning candidates are equally likely to win, but the left-leaning candidate is more likely to win unfairly. An example of such a prior is given in the following table:

μ_0	u	f
l	.75/2	.25/2
r	.25/2	.75/2

where l stands for the left-leaning candidate winning, r for the right-leaning candidate winning, u for the election being unfairly won, and f for the election being fairly won. The number in each cell corresponds to the prior likelihood of the row-column combination. After the election, data comes out: $h =$ "left-leaning candidate won the certified vote tally with no official evidence of fraud". In reality, these data are perfectly revealing that the state is (l, f): the left-leaning candidate won a fair election. The data are closed to interpretation on who won the election—there is only one possible interpretation because the winner is by definition the candidate who received the highest certified vote tally. The data, however, are open to interpretation on the election's fairness; there are many different ways to think about the implications for fairness of the lack of evidence of fraud.

Following the release of data, then, everybody agrees that the left-leaning candidate won the election, but people may disagree about whether the election was fairly won because they use different models to interpret the data. Some voters initially stick with the default interpretation that the vote tally reveals the election winner and that the election was fairly won.[1] Others, however, view the data as instead suggesting the election was unfair. Assuming the population is sufficiently large that roughly every interpretation is someone's initial reaction and that the network is sufficiently connected that the most compelling interpretation spreads throughout the population, we ask: which take goes viral?

[1]Formally, imagine $\pi_d(\text{left-leaning candidate won vote tally with no evidence of fraud}|(l, f)) = 1$ and $\pi_d(\text{left-leaning candidate won vote tally with no evidence of fraud}|\omega) = 0$ for all $\omega \neq (l, f)$.

Not the correct one. Eventually everyone will end up holding the model m^{bf} that maximizes $\Pr(h|m, \mu_0)$ subject to $\Pr(l|h, m) = 1$: that is, the best-fitting model consistent with the left-leaning candidate winning. This model implies a low probability of a fair election: $\Pr(f|h, m^{bf}) = .25$. As shown in Schwartzstein and Sunderam (2021), models that fit well tend to result in posterior beliefs close to prior beliefs. Intuitively, models that fit well imply the data is unsurprising, which means beliefs should not move much in response to it. In this example, right-leaning voters' prior is that the left-leaning candidate is unlikely to win fairly. The model that best fits the voters' knowledge (i.e., their prior and the data) perfectly confirms this prior. Thus, following social learning, right-leaning voters agree on the interpretation that the left-leaning candidate's victory is unsurprising because the election was likely unfair. This interpretation makes it as unsurprising as possible that the left-leaning candidate won. In contrast, if the data had been $h =$ "right-leaning candidate won the certified vote tally with no official evidence of fraud," these same voters would have adopted a different model. Their favored interpretation would suggest a high probability the election was fair because that interpretation would make the data as unsurprising as possible.

This example illustrates three main points. First, social learning *hardens* everyone's reaction to data that is open to interpretation: following the exchange of models, people are more convinced they have the right explanation for the data because exposure to others' models helps them find ways of explaining the data that they would not find on their own. The fit of the model voters converge to is .5, roughly 33% greater than the fit of the default model (.375=.75/2). Second, interpretations *evolve* to make data that is open to interpretation feel less surprising, which often makes final interpretations *less* accurate than initial reactions. In the example, many right-leaning voters initially have the correct reaction that the election was fairly won. However, social learning pulls their reaction back to their prior that left-leaning candidates are unlikely to fairly win. This evolution of beliefs highlights a key distinction between our formulation and those built on motivated reasoning or preferences over beliefs. In these alternative formulations, if, for example, right-leaning voters prefer accounts that a left-leaning candidate could only win by cheating, then their *initial* reactions will exhibit that preference. A third point is that social learning not only has a tendency to harden reactions but also to *mute* them—bringing posterior beliefs closer to prior beliefs—by increasing the chances that people are exposed to models that explain why the data are unsurprising and hence beliefs should not move. Put differently, the exchange of models can have a tendency to untether beliefs from data. Our analysis generalizes and fleshes out these points, asking how network structure shapes ultimate interpretations and beliefs, as well as how politicians, firm managers, and other agents can shape communication to their advantage.

Section 2 introduces the model. We say social learning hardens a person's reaction to the data when she feels she is better able to explain the data after social learning than before. It follows

straightforwardly from our assumptions that any amount of social learning hardens reactions and that the amount of hardening is increasing in the size of the network. We say that social learning mutes a person's reaction to the data when it moves the person's beliefs closer to her prior. Whether social learning mutes reactions or not depends on the network structure, as well as the degree to which data is open to interpretation.

We first establish a basic result for the case where people are willing to entertain roughly every possible interpretation of the data and everyone is exposed to everyone else's model. In this case, everyone adopts a model that perfectly explains the data, which implies there is nothing to learn from the data. Thus, social learning maximally hardens and mutes a person's reaction to the data—people end up convinced that they perfectly understand the data and that their prior beliefs are consistent with this understanding. This stylized case also captures one important feature of reality, highlighted in the voting example above: responses to data often initially diverge and then converge as people share their interpretations, and this convergence often pulls beliefs back towards views people held before seeing the data. For instance, many commentators have noted the stability of political polls in recent years.[2] Consistent with our model, this stability does not mean that polls do not react to news. They do react, but the impact of news tends to fade quickly, with people returning to their previous views. Similarly, in discussing reactions to news about Covid-19, New York Times writer Charlie Warzel recently made a similar observation: "a story comes out about a study/specific spreader event/ whatever & it's like 1) immediate intense reactions followed by 2) 36 hrs of long threads by smart & not smart/qualified & not qualified people picking apart/casting doubt & 3) usually calm consensus later in the week".[3] In our framework, this disconnect between the data and long-run beliefs is driven by the adoption of narratives through social learning.

We next turn to the impact of network structure on interpretations and beliefs. Section 3 studies networks formed on the basis of shared beliefs, where people exchange models with others who had similar initial reactions to the data. To illustrate in the voting example above, suppose voters whose initial interpretations suggested the election was unfair all talk to each other, while voters whose initial interpretations suggested the election was fair all talk to each other. We show that within each network social learning leads beliefs to converge to the initial reaction in the network that is closest to the prior. In the voting example, members of the "election was fairly decided" and the "election was unfairly decided" networks will continue to disagree, but less so over time as all right-leaning voters converge on models that bring their beliefs closer to the 25% prior probability they attached to the election being fairly decided conditional on a left-leaning candidate winning. We also show that shared belief networks can lead to polarization of beliefs across multiple issues.

[2]See, e.g., https://www.pewresearch.org/fact-tank/2020/08/24/trumps-approval-ratings-so-far-are-unusually-stable-and-deeply-partisan/.

[3]https://mobile.twitter.com/cwarzel/status/1421177475111931904

If networks are formed based on one issue (e.g., the environmental impact of genetically modified crops), exchange of interpretations leads to the convergence of within-network beliefs on a second issue (e.g., the safety of genetically-modified crops). In other words, beliefs across issues become more uni-dimensional. After social learning, beliefs about the second issue become more correlated with beliefs about the first issue.

We next study networks formed based on shared models rather than shared beliefs. For instance, communities of conspiracy theorists (e.g., proponents of QAnon) are organized around common ways of interpreting events. In finance, contrarians may communicate more with other contrarians than with trend followers. Similarly, some venture capital firms primarily evaluate start-ups based on current profitability, while others focus on management team experience. Section 4 analyzes such networks, focusing on networks based on shared inflexibility—i.e., shared dogmatism about how to interpret certain types of information. This could reflect shared expertise, shared beliefs that certain types of data are uninformative, or shared trust in some types of data. Social learning will lead groups that are inflexible in their interpretation of data h^a and but open to different interpretations of h^b to eventually come up with explanations that neutralize reactions to h^b. Communities of quantitative analysts, who are confident in how they interpret hard information and open to different interpretations of soft information, will exchange interpretations that lead them to view soft information as uninformative. If the soft information is negatively correlated with the hard information, then quantitative analysts may end up overreacting to hard information and underreacting to soft information. Thus, when networks are based on shared models, social learning may exaggerate reactions to some data in addition to hardening them.

This analysis begs a question: how can differences in beliefs across networks persist when there is some communication across networks? We draw a distinction between being weakly and strongly exposed to beliefs outside a person's network. We say a person is weakly exposed to a belief if she is aware of a single model supporting that belief, while she is strongly exposed to a belief if she is aware of all models supporting that belief. We think of communication within networks as strong exposure and communication across networks as weak exposure. Under this view, members of a network can be aware that people outside their network have different beliefs, but they will be unpersuaded by the arguments they know in favor of those different beliefs. Weak exposure to others' arguments is more effective in moving beliefs before social learning than after. By hardening reactions, social learning innoculates people against finding models supporting alternative beliefs compelling.

Armed with these results, Section 5 then considers how someone could try to manage the communication network to her advantage. A firm manager, for example, influences the network in how she forms and manages teams and in how she controls the flow of communication within her organization. Influential Twitter users shape networks in their choice of which voices to amplify by

re-tweeting. We show that if the network shaper is interested in encouraging people to take specific actions in response to data, then the shaper wants to expose people to all models that support that action. If the shaper could identify the model that resonates most with people ahead of time then she would just push that model. But if she cannot, she is better off crowdsourcing arguments and seeing what resonates instead of using one specific argument. Put differently, a network shaper who supports a particular action is better off using a collection of individuals, i.e., a platform, to articulate arguments for taking that action rather than using any given individual. The shaper also wants to prevent the audience from being exposed to certain arguments. When the shaper wants people to react to the data rather than letting the status quo prevail, she especially wants to prevent people from hearing arguments that the data are unsurprising. Such arguments will be compelling because they fit the data well and will lead people to conclude the status quo should prevail. In contrast, if the shaper cares more about the audience reaching consensus than about the specific conclusion they reach, then she wants everyone to share interpretations with each other. But this approach will favor the conclusion that there is little to learn from the data.

We then spell out some applications of our results in Section 6. Building on a large literature in organizational studies following Karl Weick (e.g., Weick (1995)) that views sensemaking as a central activity of organizations, the results on network shapers have straighforward implications for when and how firm managers want to hold meetings: If a manager's objective is to make sure workers stay on the same page, for example if there is a strong coordination motive, then she wants to have a very open flow of communication. If there is an event that is open to interpretation (e.g., someone is surprisingly denied promotion), then the manager wants to call a meeting to provide a forum for everyone to share interpretations and settle on the view that there is little to learn from the event. On the other hand, if the manager's objective is to shift workers' beliefs in response to an event, for example if she seeks to manage change, then she wants to control the flow of communication. If there is an event that is open to interpretation, she wants to call a meeting where only interpretations supporting desired conclusions are voiced. Even with such strong control, however, the manager will not be able to get everyone on the same page, perhaps shedding light on why organizations may find it difficult to reach desired shared understandings that differ from the status quo (e.g., Gibbons and Henderson (2012a)).

We also consider what the analysis implies for the evolution and spread of misconceptions through networks. Why do misconceptions, e.g., about vaccine and GM safety, persist in a world where people have access to so much high-quality information? Why do ideological bubbles appear to play a role despite the fact that people have diverse news diets and do not appear to systematically avoid counter-attitudinal information (Gentzkow and Shapiro (2011); Guess et al. (2018))? Our framework offers a simple explanation, complementing recent models that instead highlight the role of social media echo chambers (Bowen et al. (2021)): Within a bubble or network, people

are exposed to crowdsourced models that evolve to fit the data better and better, making them more compelling and resistant to change. In our framework, bubbles don't protect against being exposed to the right take on an event, but, by hardening reactions, inoculate against finding that take compelling. Vaccine skeptics are aware that many people say vaccines are safe and know some of their arguments—but they have been exposed to a broad diversity of arguments for why they are unsafe and find some such arguments more persuasive.

Related Literature

There is a large literature on social learning reviewed in Golub and Sadler (2016), with influential early contributions in economics such as Banerjee (1992), Bikhchandani et al. (1992), and Smith and Sørensen (2000). While much of this work assumes people are Bayesian in updating beliefs, important recent contributions study naive social learning by building on the simple DeGroot (1974) model of linear updating (Golub and Jackson (2010)) or on more psychologically microfounded updating rules premised on redundancy or correlation neglect (e.g., Eyster and Rabin (2010, 2014); Enke and Zimmermann (2019); DeMarzo et al. (2003); Gagnon-Bartsch and Rabin (2016)). This work focuses on people sharing information (e.g., how much they enjoyed meals at a restaurant) or observing each others' actions (e.g., seeing that a restaurant is popular), and studies questions like whether social learning successfully aggregates individuals' private information in the long run. Our focus is instead on what happens when people share the same information and exchange interpretations to make sense of that information. While frameworks featuring social learning of information tend to predict long-run consensus and relatively effective information aggregation, our framework featuring social learning of models naturally generates long-run disagreement and the persistence of false beliefs. In our framework, increasing connectedness tends to untether beliefs from data that is open to interpretation by increasing the chances of being exposed to a model that provides a compelling case that the data is unsurprising; in our framework, people adopt wrong interpretations not from hearing the same wrong interpretations repeatedly, but rather from being exposed to interpretations that evolve through social learning to compellingly fit their prior knowledge.

A smaller literature on social learning examines how people could leverage networks to their advantage in spreading information. Much of this work considers how to best seed a network with information to boost its diffusion (e.g., Akbarpour et al. (2020)). Murphy and Shleifer (2004) present a model of the creation of social networks based on shared beliefs in the context of studying political persuasion. This work considers social learning of information or beliefs rather than of models.

Closer to our work, recent presidential addresses such as Shiller (2017) and Hirshleifer (2020)

have called for studying the social transmission of narratives in economics and finance.[4] These addresses, as well as a related book (Shiller (2020)), have layed the groundwork for this study by providing vivid illustrations of the importance of socially-emergent narratives as drivers of economic and financial events. They also sketch models of narrative transmission that liken the spread of narratives to the spread of viruses. Bénabou et al. (2018) model the spread of moral narratives (e.g., "thou shall not do this because") by stratetgic actors. Our work adds to this line of study by formally modeling social forces that shape the narratives themselves and highlighting that good explanatory power helps narratives "go viral".

We build on our earlier work on model persuasion (Schwartzstein and Sunderam (2021)), which itself built on behavioral models of persuasion based on coarse or associational thinking (e.g., Mullainathan et al. (2008)). Froeb et al. (2016) present an earlier related model in the context of studying adversarial decision making in law and Aina (2021) builds on the model persuasion framework by considering what happens when persuaders need to commit to models before seeing all the data. Contemperaneous work (Eliaz and Spiegler (2020); Bénabou et al. (2018)) take somewhat different approaches to formalizing models or narratives and what makes them persuasive. For example, Eliaz and Spiegler (2020) assume that people favor "hopeful narratives". We add to this work by formalizing how social learning influences which models emerge and persist.

2 Model

2.1 Setup

The basic setup closely follows Schwartzstein and Sunderam (2021). Broadly, individual agents take the following steps in interpreting data. All agents share a common default model for interpreting data, and in addition each agent comes up with a model of their own. Prior to social learning, each agent selects from these two models the one that best explains the data. Social learning then exposes each agent to all models held by other agents in her social networks. After social learning, each agent adopts the model that best explains the data from the full set of models she has been opposed to: the default, the model she comes up with on their own, and the models others in her social network have come up with.

Formally, there are a continuum of agents $i \in [0, 1]$ who hold beliefs μ_i over states of the world ω in finite set Ω.[5] Agent i takes an action a from compact set A to maximize the expectation under μ_i of $U_i(a, \omega)$. In the baseline setup, agents share a common prior $\mu_0 \in \text{int}(\Delta(\Omega))$ over Ω and observe a public history of past outcomes, h, drawn from finite outcome space H. Agents

[4] While not all narratives are models and vice-versa, they are closely related and we sometimes interchangeably talk about narratives, stories, and models.

[5] In examples we sometimes relax the assumption that Ω is finite.

can end up with different posteriors if they use different models to interpret this history. Given state ω, the likelihood of h is given by $\pi(\cdot|\omega)$. The true model m^T is the likelihood function $\{\pi_{m^T}(\cdot|\omega)\}_{\omega\in\Omega} = \{\pi(\cdot|\omega)\}_{\omega\in\Omega}$. We assume that every history $h \in H$ has positive probability given the prior and true model.

Agents do not know the true model. A given agent updates her beliefs based on either (i) the default model $\{\pi_d(\cdot|\omega)\}_{\omega\in\Omega}$,[6] (ii) the model m_i' that she generates herself to explain the history, where m_i' is taken from compact set M and indexes a likelihood function $\left\{\pi_{m_i'}(\cdot|\omega)\right\}_{\omega\in\Omega}$, or (iii) a model she learns from someone in her social network, where we let $M_i \subseteq M$ denote the set of models proposed by someone in i's social network.

Given the history and the set of models the agent is exposed to, she adopts the one that best explains the history. Formally, let $\mu(h,\tilde{m})$ denote the posterior distribution over Ω given h and model $\tilde{m} \in M \cup \{d\}$, as derived by Bayes' rule. We assume the receiver adopts the model m and hence posterior $\mu(h,m)$ if

$$m \in \arg \max_{\tilde{m}\in\left\{d,m_i'\right\}\cup M_i} \underbrace{\Pr(h|\tilde{m},\mu_0)}_{=\int \pi_{\tilde{m}}(h|\omega)d\mu_0(\omega)} .$$

That is, the person goes with the model she is exposed to that best fits the data. We will return later to what happens in the case of ties. Upon adopting a model $\tilde{m}$, the person uses Bayes' rule to form posterior $\mu(h,\tilde{m})$ and takes an action that maximizes her expected utility given that posterior belief: $a(h,\tilde{m}) \in \arg\max_{a\in A} \mathbb{E}_{\mu(h,\tilde{m})}[U_i(a,\omega)]$.

To close the baseline model, we need to specify the model a person generates herself. Let $\bar{M}(h,\mu_0,d,M) = \{m \in M : \Pr(h|m,\mu_0) \geq \Pr(h|d,\mu_0)\}$ denote the set of models in M that explain the history as well as the person's default interpretation given her prior over states. Assume that measure δ of the population generates the default model and measure $(1-\delta)$ generates a model in $\bar{M}(h,\mu_0,d,M)$.[7] Further assume that that population is large enough that, for each model $m \in \bar{M}(h,\mu_0,d,M)$, someone in the population generates that model herself.

In the typical case, we set the default interpretation to be the true-model interpretation, $d = m^T$. We also typically let M be the set of all possible models M^a—i.e., for any likelihood function $\{\tilde{\pi}(\cdot|\omega)\}_{\omega\in\Omega}$ there is an $m \in M^a$ with $\{\pi_m(\cdot|\omega)\}_{\omega\in\Omega} = \{\tilde{\pi}(\cdot|\omega)\}_{\omega\in\Omega}$. We refer to this as the case where people are *maximally open to persuasion.* We simply write $\bar{M}(h,\mu_0)$ as shorthand for

[6]The default can potentially be a function of h. We suppress the dependence of d on h when it does not cause confusion.

[7]Alternatively, we could endogeneize δ by assuming that people sometimes generate models outside of $\bar{M}(h,\mu_0,d,M)$ in which case they stick with the default model. This would suggest that δ is larger when the default does a good job explaining the data h. While this change would influence the distribution of beliefs prior to social learning, it would not influence the distribution of beliefs following social learning.

$\bar{M}(h, \mu_0, m^T, M^a)$.[8]

2.2 Discussion of Model Assumptions

The building blocks of the model come from Schwartzstein and Sunderam (2021), and we refer to that paper for a detailed discussion of the basic assumptions. We depart from that paper in a few crucial ways.

First, we allow some receivers by themselves to generate a model other than the default. That is, in the notation of our current framework, our previous paper assumes $\delta = 1$ (receivers stick with the default before being exposed to persuasion), while the analysis in this paper focuses on the case where $\delta < 1$. For many topics, it seems plausible that some people generate an initial interpretation of the data, prior to sharing interpretations with others. Many of us have gut reactions about why the stock market moved yesterday, who is responsible for the storming of a government building, what the latest school shooting implies about the merits of gun control. These gut reactions may be constructed spontaneously in response to the data and differ across people. Crucially, however, we assume that a given person does not come up with all models she is willing to entertain, so she is influenced by which models she is exposed to.

Second, the focus of this paper's analysis is on the social exchange of models, not on the behavior of a strategic persuader who attempts to influence the beliefs and behavior of audience members. While many situations are well described by the persuasion setup, many other situations involve evolving interpretations of data through social learning. By taking as primitive the set of models a given person i is exposed to, M_i, our framework accommodates a variety of network structures, including both directed networks, where the flow of communication goes one way, and undirected networks, where it goes two ways.

Third, implicit in the idea that a person is exposed only to the models of those within her network is an assumption that she does not actively seek out the models proposed by members of other networks. One way of thinking about this assumption is that people exhibit a sort of out-group homogeneity bias (e.g., Quattrone and Jones (1980)), thinking there is not much reason to investigate the models in other networks because they are "all the same". A person who favors gun control may be aware of some arguments for why shootings suggest weaker gun control (e.g., "we need more guns in the hands of good guys") and think once she has heard one such argument she

[8]There's a technical issue that comes up when M^a is the set of all models. In this case, even assuming a continuum of individuals, the space $\bar{M}(h, \mu_0, m^T, M^a)$ may be too large to guarantee that, for every model in $\bar{M}$, there exists a person who holds that model before social learning. For readers who are concerned about such cardinality issues, we note that all our results and intuitions stated for the case of $M = M^a$ continue to hold if we instead make the following assumption on M: For every belief $\tilde{\mu}$ that is a posterior for some model in M^a given data h, prior μ_0, and default d, M includes the best-fitting model inducing that posterior as well as one worse-fitting model inducing that posterior.

has heard them all, perhaps underappreciating the diversity of these arguments.

2.3 Examples

Example 1 (Interpreting data about policy issues)**.** We now sketch two brief examples, which we will return to throughout the paper.

Our first example involves public-policy choices. Suppose the state space is binary, $\Omega = \{l, r\}$. In state $\omega = l$, a Democrat would make a better US president, and in state $\omega = r$ that a Republican would make a better US president. The prior over states is prior $\mu_0(l) = 1/2$. Further suppose that people can take three possible actions, $a \in \{L, M, R\}$,where action $a = L$ is to vote Democrat, $a = M$ is to abstain from voting, and $a = R$ is to vote Republican. Alternatively, one can think of the states as corresponding to whether some left- or right-leaning policy (e.g., involving gun control, climate change, pandemic policy) would be effective, and the actions as corresponding to supporting such policies ($a = L$, R) or the status quo ($a = M$). The payoffs U_i are such that $a = L$ is optimal if $\mu(l) \geq .75$, $a = M$ is optimal if $\mu(l) \in (.25, .75)$, and $a = R$ is optimal if $\mu(l) \leq .25$.

To give a flavor for the mechanics of the model without social learning, consider the case where there is no social exchange of models (i.e., $M_i = \varnothing$ for all i). Suppose h could take on two values, (h^l, h^r), and that the data fully reveals the state under the true model: $\pi_{m^T}(h^l|l) = \pi_{m^T}(h^r|r) = 1$. Further assume people are maximally open to persuasion, $M = M^a$, and the default model is the true model, $d = m^T$. In this case, $\Pr(h|m^T, \mu_0) = 1/2$. However, the set of beliefs people will have after interpreting the data themselves $\bar{M}(h, \mu_0)$ is the full support $\mu(l) \in [0, 1]$. Intuitively, there are many models that seem more compelling than the true model (i.e., many models under which the history is more likely than it is under the true model). These models imply a wide variety of posterior beliefs. However, by a simple application of Proposition 1 in Schwartzstein and Sunderam (2021), it is not always the case that all beliefs are supported in equilibrium. Changing the example slightly so $\mu_0(l) = .7$, then, for $h = h^l$, the set of beliefs supported by a model in $\bar{M}(h, \mu_0)$ is $\mu(l) \in [4/7, 1]$ and, for $h = h^r$, the set of beliefs supported by a model in $\bar{M}(h, \mu_0)$ is $\mu(l) \in [0, 1]$.[9]

This examples highlights three points. First, before social learning, people may have very different reactions to the same data, even if they share a prior and default interpretation. Second, people may initially react to the data in extreme ways. Indeed, in some instances, their beliefs can move in response to data in the opposite direction than they would updating under the true model. Third, as shown by the $\mu_0(l) = .7$ case, there may be a grain of truth in people's initial reactions: in that case, beliefs following a signal that indicates $\omega = l$ favor that state relative to beliefs following

[9]To derive the fraction of people who choose each action when $M_i = \varnothing$, we need to additionally specify the distribution over models people come up with on the fly. When people instead talk to each other, many of our results are independent of the choice of this distribution.

a signal that indicates $\omega = r$.

We will sometimes extend this example to consider cases where people may use the same data to update beliefs about a variety of issues. For instance, in the introduction, people updated both about who won the election and whether the election was fair. Similarly, people may construct narratives surrounding data about genetically-modified crops that both have implications for their safety and their impact on the environment (e.g., how their adoption influences pesticide use). To accommodate such examples, let $\Omega = \Omega^1 \times \Omega^2$. We will consider how network-members' beliefs over Ω^1 (e.g., what the data implies about the environmental impact of genetically-modified crops) spill over to influence beliefs over Ω^2 (e.g., what the data implies about the safety of genetically-modified crops).

Example 2 (Interpreting data about startups)**.** Our second example involves investing. Relative to the first example, there are two qualitative differences. First, it highlights more clearly the role of the data. Second, it illustrates how restricting the model space impacts our results.

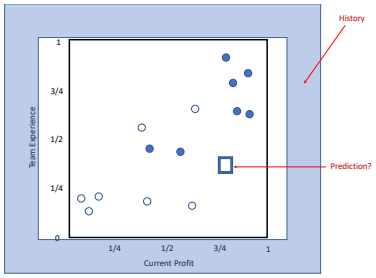
Consider a community of venture capitalists trying to predict the success of a startup in a new sector (e.g., cryptocurrency) based on the history of past startups and their characteristics. The history of past startups is $h = \{(x_{1j}, x_{2j}, x_{3j}, y_j)\}_j$ where $y_j = 1$ if startup j succeeded and $y_j = 0$ if it failed. The characteristics of startup j are its profits (x_{1j}), management-team experience (x_{2j}), and an individuating characteristic (x_{3j}) – a characteristic that is unique to each startup. Figure 1a shows an example history. Each dot represents a previous startup, with profit plotted on the horizontal axis and team experience plotted on the vertical axis. The individuating characteristics are not pictured. A dot is filled in if the startup was successful and is unfilled if it failed. Venture capitalists start with a prior that a given startup's probability of success, θ, is uniformly distributed on $[0, 1]$ and dogmatically believe that (profit) x (experience) characteristics are uniformly distributed in $[0, 1] \times [0, 1]$. They then use the history to make predictions about a new startup k's success probability as a function of its characteristics.

We assume there are four types of models in the model space M. First, all venture capitalists start with the default model that all startups in the new sector have the same success probability regardless of their characteristics. Second, there are models that are cutoff rules in profit: all startups with profit below the cutoff share the same success probability and all startups with profit above the cutoff share the same success probability.[10] For instance, the vertical green line in Figure 1b depicts the model where the cutoff is the 25th percentile of profits. Third, there are models that are cutoff rules in team experience: all startups with experience below the cutoff share the same success probability and all startups with team experience above the cutoff share the same success probability. For instance, the horizontal red line in Figure 1c depicts the model where the cutoff

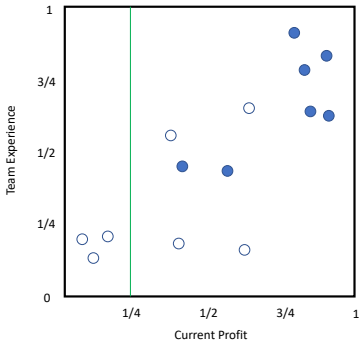
[10]Formally, success probabilities below and above the cutoff are independently drawn—once and for all—from the uniform distribution.

is the 25th percentile of experience. Fourth, there is a model positing that neither profits nor experience matter. Instead, each startup's outcome is due to its individuating characteristics; in other words, each startup had a unique feature that perfectly determined success or failure. Note that this model perfectly explains each data point. Formally, under the model m^{ind}, $\Pr(y|x_3, m^{ind}, \mu_0) = 1$ for $y \equiv (y_j)_j$ and $x_3 \equiv (x_{3j})_j$.

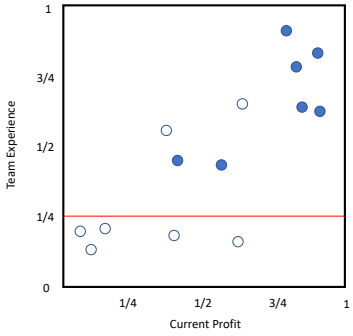
Prior to any social learning, venture capitalists consider the default and one other model randomly selected from the other three model types so long as it fits better than the default. As shown in Figure 1d, venture capitalists will have a variety of different interpretations, and thus different beliefs, at this point. In the figure, we depict for simplicity the case where the cutoffs considered are at the 25th, 50th, and 75th percentiles of each dimension. All fit better than the default.



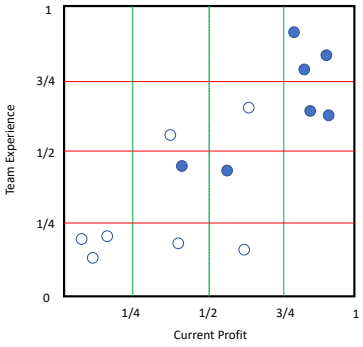
(a) Setup



(b) Model emphasizing current profits



(c) Model emphasizing experience



(d) Initial models

Figure 1: Predicting the success of a startup

2.4 Basic Definitions and Results

Prior to social learning, a person adopts the model

$$m' \in \arg \max_{\tilde{m} \in \{d, m_i'\}} \Pr(h|\tilde{m}, \mu_0)$$

and holds beliefs $\mu(h, m')$. Following social learning, the person adopts the model

$$m \in \arg \max_{\tilde{m} \in \left\{d, m_i'\right\} \cup M_i} \Pr(h|\tilde{m}, \mu_0)$$

and holds beliefs $\mu(h, m)$. As shorthand, write μ_i' (m_i') as person i's beliefs (adopted model) prior to social learning and μ_i (m_i) as her beliefs (adopted model) following social learning.

We say that social learning *hardens* a person's reaction to data when she feels she can better explain the data following social learning than before social learning: that is, when $\Pr(h|m_i, \mu_0) \geq \Pr(h|m_i', \mu_0)$. When social learning does not harden the person's reaction, we say it *softens* the person's reaction. We say that social learning *mutes* a person's reaction to data when it moves the person's beliefs closer to her prior: that is, defining Movement$(\tilde{\mu}; \mu_0) \equiv \max_{\omega \in \Omega} \tilde{\mu}(\omega)/\mu_0(\omega)$ as in Schwartzstein and Sunderam (2021), Movement$(\mu_i; \mu_0) \leq$ Movement$(\mu_i'; \mu_0)$. When social learning does not mute a person's reaction to data, we say it *intensifies* the person's reaction.

A simple observation is that social learning must harden a person's reaction to data: being exposed to more explanations of the data enables the person to better explain the data. Social learning leads a person to become more convinced she understands why the market moved as it did, why an unexpected political event occured, or the daily movement in pandemic deaths. Following social learning, any event seems more predictable.

Whether beliefs are muted by social learning depends on details of the network structure. We first establish a result for the case where every person is exposed to everyone else's model: that is, where the network is "complete" in graph-theory parlance. (All proofs are in Appendix A.)

Proposition 1. *Suppose everyone is maximally open to persuasion and talks to every person: $M = M^a$ and $M_i = M^a$ for all i. Then social learning mutes every person's reaction to the data: for every person i, Movement$(\mu_i; \mu_0) \leq$ Movement$(\mu_i'; \mu_0)$. In fact, social learning maximally mutes and hardens each person's reaction to the data in the sense that each person sticks with their prior belief with a model that perfectly explains the data: for every person i, $\mu_i = \mu_0$ and $\Pr(h|m_i, \mu_0) = 1$.*

This result says that if everyone talks to each other, then following social learning they do not react to the data at all because they view it as inevitable in hindsight. Given that the world is large, someone will come up with the model that explains why whatever happened was bound to happen; that model will spread throughout the network, and since the network connects everyone, everyone will adopt it. The fact that a model that fits the data well will be broadly adopted in turn means that beliefs will move very little. Intuitively, models that fit well imply the data is unsurprising, which means beliefs should not move much in response to it.

In the context of the binary-state policy example above, this result says that a given person's

reaction to an event might push her to favor taking actions $a = L$ or R, but after being exposed to many different arguments she will go back to favoring the status quo of $a = M$. As an illustration, a school shooting might initially lead people to think we need a change in gun-control policies, but they will eventually favor interpretations that say we did not learn much from the shooting. Indeed, there is empirical evidence of exactly such dynamics. Following mass shootings, Twitter users who are initially against gun control temporarily become more open to the idea. However, as narratives evolve and spread in the weeks following a mass shooting, these Twitter users slowly revert back towards their original beliefs (Lin and Chung (2020)). As noted in the introduction, such a pattern is harder to explain in mechanical models of motivated beliefs, which would instead suggest that people who are initially against gun control would immediately (i.e., before social learning) come up with ways to view the mass shooting as confirming prior beliefs against gun control.

More generally, our model predicts that following a realization of new data that is open to interpretation, there is initially a broad divergence of opinion followed by convergence as people share their interpretations and settle on commonly believing they learned little from the data. Models evolve though social learning to better and better fit the data, which in turn lead people's beliefs to move less and less.

While Proposition 1 considers the impact of social learning when *everyone* talks to each other, people are often embedded in smaller networks. We next turn to establish general comparative statics before studying specific types of networks in the following sections.

2.5 Comparative Statics

We now consider what happens if we enlarge the size of a network prior to social learning, increasing the set of models network members are exposed to. We primarily consider expanding the network in three specific ways. First, we consider *merging networks*: expanding person i's network by merging it with person j's network enlarges the set of models that are shared with person i to $M_i \cup M_j$. This exercise helps assess the impact of increasing social connectedness. Second, we consider *weakly exposing members to an alternative belief*: expanding person i's network by weakly exposing her to an alternative belief $\tilde{\mu}$ enlarges the set of models that are shared with person i to $M_i \cup \{m(\tilde{\mu})\}$, where $m(\tilde{\mu})$ is a specific model that supports belief $\tilde{\mu}$. Third, we consider *strongly exposing members to an alternative belief*: expanding person i's network by strongly exposing her to an alternative belief $\tilde{\mu}$ enlarges the set of models that are shared with person i to $M_i \cup M(\tilde{\mu})$, where $M(\tilde{\mu})$ is the set of all models that induce $\tilde{\mu}$. These latter two exercises help assess the impact of getting people out of bubbles or echo chambers.

Proposition 2. *Suppose everyone is maximally open to persuasion, $M = M^a$. Let μ_i (m_i) denote a person's belief (model) following social learning prior to a network expansion, and μ_i^e (m_i^e) denote*

her belief (model) following social learning with the expanded network.

1. *Expanding person i's network in any way weakly hardens her reaction to the data: for any expansion of M_i to $M_i \cup \tilde{M}$ with $\tilde{M} \subset M$, $\Pr(h|m_i^e, \mu_0) \geq \Pr(h|m_i, \mu_0)$.*

2. *For every belief $\tilde{\mu}$, there exists a model $m(\tilde{\mu})$ supporting that belief that is less compelling than the model m_i the person would adopt prior to a network expansion. Thus, for every $\tilde{\mu}$ it is possible that weakly exposing network members to that belief has no impact on their beliefs following social learning.*

3. *If expanding person i's network by weakly exposing her to an alternative belief $\tilde{\mu}$ impacts her final belief, $\mu_i^e \neq \mu_i$, then expanding her network by strongly exposing her to $\tilde{\mu}$ impacts her final belief. However, the converse does not hold.*

Proposition 2 shows that expanding a network always (weakly) hardens a network member's beliefs. Being exposed to more models leads a person to become more convinced she knows how to interpret the data that is open to interpretation. The most basic impact of increasing connectedness in our model is increasing a person's view that such data was predictable.

We also see from Proposition 2 that expanding a person's network by strongly exposing her to an alternative belief is more impactful on her ultimate beliefs and behavior than weakly exposing her to an alternative belief. Our framework suggests that a platform of different contributors (say Breitbart) will tend to be more influential than any single contributor (say Steve Bannon), even if that contributor reaches the same audience. The reason is that increasing a person's exposure to a broad range of arguments supporting a given conclusion makes it more likely that she will find one of those arguments compelling than if she is exposed to only a few of those arguments.

Another way of thinking about this basic result connects to what it means to get someone out of an ideological bubble. Strongly exposing a person to a belief could be thought of as getting her outside of a bubble by expanding her network to include members with that belief, while weakly exposing a person to a belief could be thought of as making her aware that someone outside of her network holds that belief. Under this interpretation, Proposition 2 suggests that the former is more effective at changing minds because it increases the diversity of arguments a person is exposed to that supports a belief.[11] We return to this discussion in Section 6.

Foreshadowing something we will see in the next section, the difference between strong and weak exposure also sheds light on why people in different networks may settle on different beliefs, even if they sometimes communicate across networks. People might exchange models in addition

[11] In highlighting the importance of the breadth of arguments a person is exposed to in whether this changes her mind, our model relates to "persuasive-arguments theory" from psychology (e.g., Burnstein and Vinokur (1977)). However, persuasive-arguments theory emphasizes the number of distinct arguments a person is exposed to, while we emphasize the compellingness of arguments (in terms of fit) a person is exposed to.

to beliefs when interacting with others in the same network, while only exchanging beliefs (and perhaps a subset of models supporting those beliefs) when interacting with members of different networks. A person who believes a school shooting indicates the need for stricter gun-control measures is likely aware that there are others who conclude the opposite without being intimately familiar with their arguments. While a person might become convinced by listening to a broad set of arguments, she is unlikely to be convinced by the positions themselves.[12]

We next turn to studying specific networks.

3 Shared Belief Networks

Many networks are formed based on shared beliefs. For instance, networks are formed based on beliefs that one political party typically governs better than others, that vaccines are harmful, and that free markets lead to prosperous societies.

To analyze such networks, suppose that the beliefs a person holds prior to talking to others influences who she talks to. Formally, consider a partition $\mathcal{S}$ over the set of beliefs $\Delta(\Omega)$, where we denote $s(\mu)$ as the element in $\mathcal{S}$ that belief $\mu \in \Delta(\Omega)$ belongs in. In a *shared-belief network*, a person i exchanges models with another person j if and only if their initial beliefs are similar, in the sense that they fall in the same element of $\mathcal{S}$.

Definition 1. In a *shared-belief network*, $M_i = \left\{m \in \bar{M}(h, \mu_0, d, M) : \mu(h, m) \in s(\mu(h, m_i'))\right\}$ for every person i.

Given our assumption of common priors, this definition, taken literally, says that a shared-belief network forms based on a common reaction to a specific event. For example, a shared-belief network could form among people who react similarly to a police shooting in their beliefs on the need for police reform. While we think this literal interpretation is a reasonable approximation of reality for certain high-profile events, in many instances networks based on shared beliefs are probably constructed based on common reactions to a broader set of events. For example, people who tend to lean left in their interpretations might share views on the most recent event, even if their initial views on that most recent event are quite different. In such cases, a broader interpretation of our shared-beliefs setup is appropriate—these networks are formed among people who initially hold similar beliefs about some question of interest, whether or not these similar beliefs arise literally from having a common initial reaction to the most recent event.[13]

[12] A different reason why beliefs might not converge across networks is that, after engaging in social learning within a network, a person's posteriors may become her priors for evaluating arguments outside the network—i.e., she might evaluate model m based on $\Pr(h|m, \mu(h, m_i))$, rather than $\Pr(h|m, \mu_0)$. This is a sort of confirmation bias, which would advantage models supporting beliefs close to $\mu(h, m_i)$. We explore a related possibility when briefly considering dynamics in Section 6.

[13] We will more formally capture this idea in briefly studying dynamics in Section 6.

Before establishing a basic result for shared-belief networks, we recall a lemma from Schwartzstein and Sunderam (2021).

Lemma 1 (Schwartzstein and Sunderam (2021))**.** *Fix history h and let*

$$\textit{Fit}(\tilde{\mu}; h, \mu_0) \equiv \max_m \Pr(h|m, \mu_0) \textit{ such that } \mu(h, m) = \tilde{\mu}$$

be the maximal fit of any model that induces posterior $\tilde{\mu}$ given the history h and a person's prior μ_0. Then

$$\textit{Fit}(\tilde{\mu}; h, \mu_0) = 1/\textit{Movement}(\tilde{\mu}; \mu_0).$$

The idea behind this inverse relationship between fit and movement is that models that fit the history well say it is unsurprising in hindsight, which then implies that beliefs should move little. So, for any given belief μ, the maximal fit of a model inducing that belief is greater the closer this belief is to μ_0.

Proposition 3. *Suppose everyone is in a shared-belief network and is maximally open to persuasion, $M = M^a$. Then social learning mutes every person's reaction to the data: for every person i, Movement$(\mu_i; \mu_0) \leq$ Movement$(\mu_i'; \mu_0)$. In fact, social learning leads everyone to share the initial belief within their network that is closest to the prior: for every person i, $\mu_i \in \arg\min_{\mu \in s(\mu_i')}$ Movement$(\mu; \mu_0)$.*

This result says that a person who only exchanges models with others who react similarly to data ends up at a belief that reacts least to the data among those that are shared with her.[14] By the earlier lemma, such a belief is supported by a better-fitting model than any other she is exposed to.

3.1 Comparative Statics

Next consider comparative statics.

Proposition 4. *Suppose everyone is maximally open to persuasion, $M = M^a$, and is in a shared-belief network. Let μ_i (m_i) denote a person's belief (model) following social learning prior to a network expansion, and μ_i^e (m_i^e) denote her belief (model) following social learning after a network expansion.*

[14] Since everyone within a network ends up sharing the same beliefs, the solution concept we apply to shared-belief networks is a refinement of an alternative concept proposed by Murphy and Shleifer (2004) that requires members of the same network to hold sufficiently close *post* social-learning beliefs. If we applied the Murphy and Shleifer concept, we would have many equilibria. As an illustration, in the binary-state example with a 50-50 prior, taking any finite set of final beliefs $\{\mu(l)^1, \mu(l)^2, \ldots, \mu(l)^K, 1/2\}$ with $\mu(l)^1 < \mu(l)^2 < \ldots < \mu(l)^K < 1/2$ there is an equilibrium where these are the final beliefs of members of the $K+1 \in \{1, 2, \ldots, \}$ groups: these would be the final beliefs if people with initial beliefs $[0, \mu(l)^1]$ were in one network, those with initial beliefs $(\mu(l)^1, \mu(l)^2]$ were in another, ..., and those with intitial beliefs $(\mu(l)^K, 1]$ were in a network.

1. *Expanding person i's network in any way weakly mutes and hardens her reaction to the data: for any expansion of M_i to $M_i \cup \tilde{M}$ with $\tilde{M} \subset M$, $\Pr(h|m_i^e, \mu_0) \geq \Pr(h|m_i, \mu_0)$ and Movement$(\mu_i^e; \mu_0) \leq$ Movement$(\mu_i; \mu_0)$.*

2. *Expanding person i's network by merging it with person j's leads person i to hold the initial belief within her expanded network that is closest to the prior: $\mu_i^e \in \arg\min_{\mu \in s(\mu_i') \cup s(\mu_j')}$ Movement$(\mu; \mu_0)$.*

3. *Expanding person i's network by weakly exposing her to an alternative belief $\tilde{\mu}$ impacts her final beliefs only if $\tilde{\mu}$ is closer to the prior than person i's final beliefs without the expansion: that is, $\mu_i^e \neq \mu_i$ only if Movement$(\tilde{\mu}; \mu_0) \leq$ Movement$(\mu_i; \mu_0)$.*

4. *Expanding person i's network by strongly exposing her to an alternative belief $\tilde{\mu}$ impacts her final beliefs if and only if $\tilde{\mu}$ is closer to the prior than i's final beliefs without the expansion: that is, $\mu_i^e \neq \mu_i$ if and only if Movement$(\tilde{\mu}; \mu_0) \leq$ Movement$(\mu_i; \mu_0)$.*

Proposition 4 shows that when networks are based on shared beliefs, expanding the network always (weakly) hardens and mutes a network-member's beliefs. Being exposed to more models leads a person to become more convinced she knows how to interpret the data, while adopting a model that mutes her reaction to the data. In the limit where the person is exposed to all models, we saw from Proposition 1 that the person will adopt a model where the data is completely neutralized: when data is open-to-interpretation *and relevant* for updating beliefs about ω under the true model, further expanding a person's shared-belief network further untethers her beliefs from reality.

We also see more specific instantiations of the more general results above on the impact of weakly and strongly exposing people to alternative beliefs. When networks are formed based on shared beliefs, a person will only be swayed by a belief not initially represented in her network if the belief responds less to the open-to-interpretation data (i.e., is closer to the prior) than the belief she would otherwise have settled on: Beliefs tend to be more attractive when they are closer to a person's prior. But, again, strongly exposing a person to such a belief is more likely to move her beliefs than weakly exposing her to such a belief.

Exposure to alternative beliefs is also more effective when it comes prior to social learning. Imagine that before joining a shared-belief network, person i with belief μ_i' is weakly exposed to belief $\tilde{\mu} \notin s(\mu_i')$ with supporting model $m(\tilde{\mu})$. Following exposure to model $m(\tilde{\mu})$, the person potentially updates her beliefs and joins the shared-belief network associated with her posterior.

Proposition 5. *Suppose everyone is maximally open to persuasion, $M = M^a$, and is in a shared-belief network. Let μ_i denote a person's belief following social learning without being exposed to a belief $\tilde{\mu} \notin s(\mu_i')$, μ_i^e denote her belief following social learning after being exposed to belief $\tilde{\mu}$ through network expansion, and μ_i^p denote her belief following social learning when exposed to belief $\tilde{\mu}$ before social learning. If being weakly (strongly) exposed to a belief through network*

expansion impacts person i's final beliefs, $\mu_i^e \neq \mu_i$, then being weakly (strongly) exposed to a belief prior to social learning impacts person i's final beliefs, $\mu_i^p \neq \mu_i$. However, the converse does not hold.

This result says that exposing someone to an alternative belief is more likely to have an impact on her final beliefs if this exposure comes before the person exchanges models with others in a shared-belief network. The reason is simple: as we saw before, social learning hardens a person's reaction to data. As we see from this result, such hardening innoculates the person against finding models supporting alternative beliefs compelling.

3.2 Examples

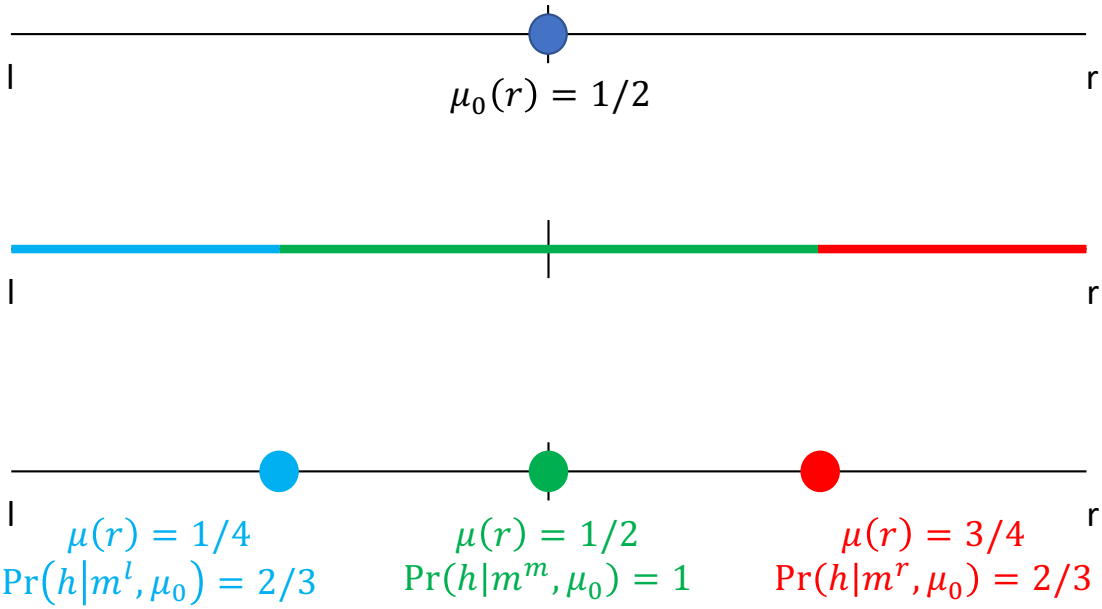
Interpreting Data About Policy Issues

As an illustration of Proposition 3, suppose that in the binary-state policy example, $\mu_0(l) = 1/2$ and shared-belief networks are formed based on people's views on the right action to take: Everyone with an initial reaction supporting a right-leaning action ($\mu_j'(r) \in [.75, 1]$) is in one network, everyone with an initial reaction supporting the neutral action ($\mu_j'(r) \in (.25, .75)$) is in another, and everyone with an initial reaction supporting the left-leaning action ($\mu_j'(r) \in [0, .25]$) is in the final network. Then Proposition 3 says, and Figure 2 illustrates, that everyone in a given network will end up at the belief that is closest to the prior within her network. For example, someone whose initial reaction to the data moves her belief from $\mu_0(r) = .5$ to $\mu'(r) = .9$ will exchange models with others whose initial reactions support the right-leaning action (pictured in red in the figure), which will end up muting her reaction to $\mu(r) = .75$. The person's reaction will also be hardened: everyone in the right-leaning network ends up adopting a best-fitting model supporting the right-leaning action, namely a model satisfying $\Pr(h|m^r, \mu_0) = 2/3$ (by application of Lemma 1).

As we saw in the special case where a person talks to everyone, talking to those who share similar beliefs is a moderating force in terms of how people think about implications of data, but a polarizing force in terms of hardening people's positions by exposing them to better-fitting arguments supporting those positions. In other words, people's final beliefs move less than their initial reactions, but they become more certain that their interpretation of the data is correct and perhaps more puzzled that anyone could conclude something different. The dynamics highlighted by the proposition may help explain the recent stability of political polls.[15] If voters are exchanging interpretations of data within shared belief networks, their beliefs will not respond much to that data.

[15] E.g., https://www.pewresearch.org/fact-tank/2020/08/24/trumps-approval-ratings-so-far-are-unusually-stable-and-deeply-partisan/

Figure 2: Evolution of Beliefs Across Shared-Belief Networks Surrounding a Single Policy Issue



In addition, Proposition 3 illustrates that, within any given network, social learning leads beliefs to converge. That is, beliefs within any network become more homogeneous. However, across networks, beliefs remain divergent with everyone becoming more confident in their reaction than before social learning. A person who only talks to others who share the reaction that the latest school shooting indicates the need for stricter gun-control measures will become more confident in the rationale for drawing this conclusion from the data; a person who only talks to others who share the reaction that the shooting indicates the need for looser gun-control measures will similarly become more confident in drawing this conclusion from the data.[16]

Proposition 3 also has implications for political polarization. To illustrate, we consider an extension of the example where there are multiple issues, but networks are formed based on shared beliefs about one of them. Let $\Omega = \Omega^1 \times \Omega^2$ and describe marginal beliefs over Ω^j by μ^j. Then networks are formed based on shared beliefs over issue 1 but not issue 2 when $s(\mu)$ depends only on μ^1.

In particular, let $\Omega^1 = \{l, r\}$ be whether a left- or right-leaning candidate governs better and

[16]This is in a sense consistent with the evidence in Schkade et al. (2007), which found that, after group interactions, views on climate change, affirmative action, and civil unions became more homogenous and more confident. Some studies on such "group polarization" find that beliefs also become "more extreme" after group interactions. Proposition 3 is consistent with those findings insofar as extremity is measured by confidence and inconsistent with those findings insofar as groups are formed based on shared beliefs, people have common priors, and extremity is measured by the degree to which beliefs are reactive to shared data. On this last point, Roux and Sobel (2015) shows how group polarization naturally arises in models of rational information aggregation.

$\Omega^2 = \{n, y\}$ could be whether we are (y) or are not (n) in a sort of crisis that requires the expertise of scientists. Networks are formed given beliefs over $\{l, r\}$ but not $\{n, y\}$: suppose people with initial beliefs $\mu_i^{1'}(l) \geq .75$ are in one network (the "left-leaning network"), those with initial beliefs $\mu_i^{1'}(l) \in (.25, .75)$ are in another (the "centrist network"), and those with initial beliefs $\mu_i^{1'}(l) \leq .25$ are in another (the "right-leaning network").

Even though beliefs over the second issue do not influence network formation, final beliefs over that issue differ across networks when prior beliefs are correlated across the issues. For example, people might believe that left-leaning candidates tend to govern better at times when a crisis requires the expertise of scientists:

μ_0	n	y
l	.25/2	.75/2
r	.75/2	.25/2

.

In this case, the movement-minimizing belief among members of the left-leaning network is

$\mu^{\text{left-leaning}}$	n	y
l	$3/4 \cdot .25$	$3/4 \cdot .75$
r	$1/4 \cdot .75$	$1/4 \cdot .25$

,

the movement-minimizing belief among members of the centrist network is the prior, while the movement-minimizing belief among members of the right-leaning network is:

$\mu^{\text{right-leaning}}$	n	y
l	$1/4 \cdot .25$	$1/4 \cdot .75$
r	$3/4 \cdot .75$	$3/4 \cdot .25$

.

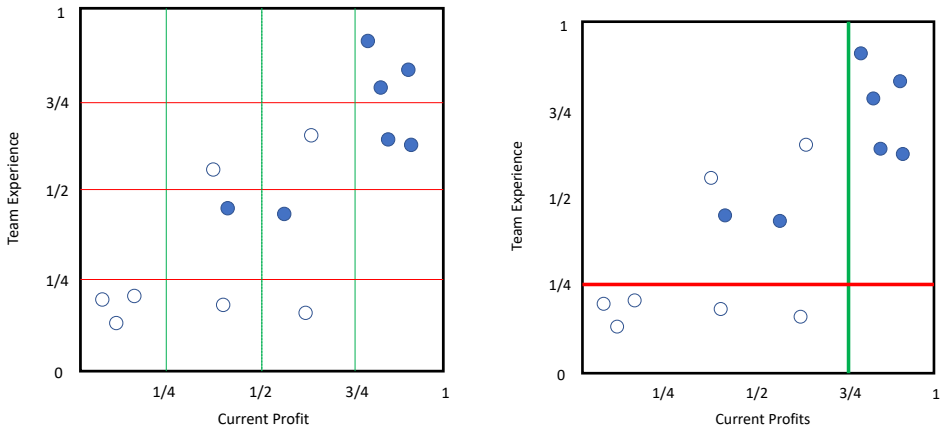
By Proposition 3, this says that $\mu^{\text{left-leaning}}$ is the shared final belief among members of the left-leaning network, μ_0 is the shared final belief among members of the centrist network, and $\mu^{\text{right-leaning}}$is the shared final belief among members of the right-leaning network. While members of the left-leaning network will view data as suggesting the likelihood of a crisis is $\mu^{\text{left-leaning}}(y) = 3/4 \cdot .75 + 1/4 \cdot .25 = .625$, members of the right-leaning network will view this same data as suggesting the likelihood of a crisis is $\mu^{\text{right-leaning}}(y) = 1/4 \cdot .75 + 3/4 \cdot .25 = .375$. Sharing models that suggest the left-leaning candidate is better at governing leads members of the network to also interpret the data as suggesting that it is likely there is a crisis that requires the expertise of scientists. Conversely, sharing models that suggest the right-leaning candidate is better at governing leads members of the network to also interpret the same data as suggesting that it is unlikely there is a crisis. In other words, in this example, the belief that there is a crisis becomes a "spurious justification" for the belief that the left-leaning candidate will govern better. Agents who interpret

the data as supporting the left-leaning candidate will diagnose the data along other dimensions in a way that justifies that candidate.

These results illustrate how networks based on one issue shape views on connected issues. This perhaps shed light on the so-called "polarization of reality" documented by Alesina et al. (2020). They show how the political left and right differ in their perceptions of factual issues, for example on the probability of upward social mobility.

Interpreting Data About Startups

Next return to the startup example and suppose venture capitalists are in shared-belief network. Specifically, they share interpretations with others who have similar initial reactions to the data. Optimists who believe the data suggest that the average startup is likely to be successful talk to each other; pessimists who believe the data suggest that the average startup is likely to be *un*successful talk to each other; and moderates who believe that success of the average startup is 50-50 talk to each other. This network structure may emerge because people with different initial reactions have different objectives going forward. For instance, optimists think they are likely to invest and want to figure out the characteristics that matter most for success, while pessimists want to figure out the most compelling way to explain to their clients why they are not investing.



(a) Initial models (b) Models after social learning

Figure 3: Evolution of Beliefs Across Shared-Belief Networks Surrounding Startup Success

Social learning will lead beliefs to converge within each network to the model that best fits the data within that network. For instance, consider the optimists. Two models lead to optimistic interpretations of the data: one where the cutoff is at the 25th percentile of experience and one where the cutoff is at the 25th percentile of profits. The former fits the data almost ten times

better than the latter. This can be seen in by comparing Figures 1b and 1c. The experience-based model in Figure 1c more effectively separates successes from failures than does the profit-based model in Figure 1b.[17] Thus, after social learning, all optimists adopt the experience-based model, depicted by thick-red horizontal line in Figure 3b. Given the data and this adopted model, simple application of the standard beta-binomial updating formula tells us that members in the optimist network forecast average startup success to be $3/4 \cdot ((7+1)/(9+2)) + 1/4 \cdot (1/(5+2)) \approx .58$. Essentially, they believe the best way to explain the data is that failure is relatively rare–only the startups with the least experienced management teams fail. This example illustrates that with a limited model space, muting does not always occur. In this case, optimists' beliefs are more optimistic after social learning than before. However, we show in simulations (available upon request) that even in this example social learning mutes the data on average (i.e., across different draws of startup characteristics).

Members of the pessimist network go through an evolution similar to the optimists. There are two models that lead to pessimistic interpretations: one with a cutoff at the 75th percentile of experience and one with a cutoff at the 75th percentile of profits. In this case, the profit-based model fits approximately ten times better than the experienced-based model, so pessimists' converge to the model depicted by the thick-green vertical line in Figure 3b. Given the data, members in the pessimist network forecast average startup success to be $3/4 \cdot ((2+1)/(9+2)) + 1/4 \cdot ((5+1)/(5+2)) \approx .42$.

Finally, consider the neutral network. Prior to social learning, the two models in the neutral network are the default model (that the success probability is the same regardless of characteristics) and the model where the success or failure of each previous startup was inevitable given individuating characteristics. The latter model fits the data perfectly, so members of the neutral network converge to it, while continuing to forecast average startup success to be $.5$.

This example illustrates why beliefs need not converge across networks even when people are weakly exposed to beliefs outside their networks. If members of the optimistic or pessimistic network were exposed to the models members of the neutral-network converge to, they would find those models compelling since they perfectly explain each data point. However, if they instead are only weakly exposed to the conclusions of members of the neutral network (that success of each startup is a 50-50 coin flip), they may associate those beliefs with the default model since it also generates those conclusions. Given that the default model does not fit as well as the models optimists and pessimists have converged to, neutral-network beliefs could in fact be deeply puzzling to members of other networks.

[17] Formally, the likelihood of the data under the experience-based model is proportional to $(\int_0^1 (1-\theta)^5 d\theta) \cdot (\int_0^1 \theta^7 (1-\theta)^2 d\theta) \approx .00046$, while the likelihood of the data under the profit-based model is proportional to $(\int_0^1 (1-\theta)^3 d\theta) \cdot (\int_0^1 \theta^7 (1-\theta)^4) \approx .000063$.

The example also highlights how interpretations may evolve in very different ways across networks. Members of the optimist network come to believe that startup success is predicted by experience and not profit, members of the pessimist network instead come to believe that startup success is predicted by profits and not experience, and members of the neutral network come to believe that success is unpredictable *ex ante* because individuating characteristics are all that matter. Thus, details of the network structure—i.e., who talks to whom—influence what people in each network will end up concluding. We next turn to studying networks structured around shared models rather than shared beliefs.

4 Shared Model Networks

Some networks are based not on shared beliefs, but shared models. Astrologers consider the movement of celestial bodies in making sense of what happened yesterday. Closer to earth, some communities of venture capitalists primarily evaluate startups based on attributes of their products, while others focus on attributes of their founders. In finance, there are contrarians and trend followers. Some political analysts focus on fundamentals (e.g., the economy) in predicting who will win an election, while others focus on polls. How do networks shape views in such cases?

To analyze shared-model networks, consider a partition $\mathcal{C}$ over the set of admissable models M, where we denote $c(m)$ as the element in $\mathcal{M}$ that model $m \in M$ belongs in. In a *shared-model network*, a person i exchanges models with another person j if and only if their initial models are similar, in the sense that they fall in the same element of $\mathcal{M}$.

Definition 2. In a *shared-model network*, $M_i = \left\{m \in \bar{M}(h, \mu_0, d, M) : m \in c(m_i')\right\}$ for every person i.

People in a given shared model network will end up agreeing on whichever model in $c(m)$ maximizes $\Pr(h|\cdot, \mu_0)$.

We analyze a special class of shared models based on shared inflexibility: people may be commonly dogmatic on how to interpret certain types of information. This may arise from shared expertise, shared beliefs about what sort of data is uninformative, shared trust in taking some data at face value, or even a shared convention that some discussions are taboo.

Decompose h into two types of data, h^a and h^b. In predicting the success of a project, stock, or politician, for example, there may be both quantitative or hard information, as well as qualitative or soft information. In interpreting whether a left- or right-leaning policy is better, there may be data communicated by left-leaning and right-leaning outlets.

Imagine there are networks that view h^a as open to interpretation, but not h^b, and vice-versa. Quantitative analysts may believe they have a good handle on how to interpret hard information

but may be more open to different ways of thinking about qualitative information. Symmetrically, qualitative analysts may have a single interpretation of soft interpretations but be open to many interpretations of hard information. People on the left may believe they know how to interpret left-leaning information, e.g., as trustworthy, but may be less sure on how to interpret right-leaning information. More formally, suppose there are three categories of models:

$$
\begin{aligned}
c^A &= \left\{m \in \bar{M}(h, \mu_0, d, M) : \pi_m(h^a, h^b|\omega) = \pi_m(h^b|\omega) \cdot \pi_{m^{fa}}(h^a|\omega) \ \forall\ \omega \in \Omega\right\} \\
c^B &= \left\{m \in \bar{M}(h, \mu_0, d, M) : \pi_m(h^a, h^b|\omega) = \pi_m(h^a|\omega) \cdot \pi_{m^{fb}}(h^b|\omega) \ \forall\ \omega \in \Omega\right\} \\
c^O &= \bar{M}(h, \mu_0, d, M) \setminus \left\{c^A, c^B\right\}.
\end{aligned}
$$

The first category of models, c^A, has a fixed interpretation m^{fa} of h^a but differing interpretations of h^b. Conversely, category c^B has a fixed interpretation m^{fb} of h^b but differing interpretations of h^a. Finally, category c^O contains all other models. If shared inflexibility stems from shared expertise, it is natural to assume $m^{fa} = m^T$ and $m^{fb} = m^T$; if it stems from shared beliefs that the data is uninformative, it is natural to assume that m^{fa} renders h^a uninformative and m^{fb} renders h^b uninformative; if it stems from shared trust in knowing the process, it's natural to assume $m^{fa} = d$ and $m^{fb} = d$.

Supposing the data is maximally open to persuasion, $M = M^a$, then people with initial models in c^A will end up convincing themselves that h^b is obvious in hindsight and hence uninformative, while people with initial models in c^B will end up analogously convincing themselves that h^a is uninformative. Quantitative analysts will talk to other quantitative analysts about how to interpret qualitative information and end up agreeing that, while it initially seemed relevant, it is not useful. Conversely, qualitative analysts will talk to other qualitative analysts about how to interpret quantitative information and end up agreeing that, while it initially seemed relevant, it is not useful. Similarly, people on the left will end up adopting models that neutralize data communicated by right-leaning outlets as being inevitable no matter the state, and similarly for people on the right.

Proposition 6. *Suppose everyone is maximally open to persuasion, $M = M^a$, and is in a shared-model network based on shared inflexibility of the form described above, where $c(m) \in \left\{c^A, c^B, c^O\right\}$. Then social learning need not moderate everyone's reaction to the data. In particular, social learning leads members of c^A to view h^b as uninformative, members of c^B to view h^a as uninformative, and members of c^O to view h as uninformative, resulting in final beliefs:*

$$
\mu_i = \begin{cases} \mu(h^a, m^{fa}) & \text{if } m_i' \in c^A \\ \mu(h^b, m^{fb}) & \text{if } m_i' \in c^B \\ \mu_0 & \text{if } m_i' \in c^O. \end{cases}
$$

As an illustration, consider networks based on shared expertise and imagine a company will either be successful ($\omega = 1$) or unsuccessful ($\omega = 0$) with equal probability ex ante. People are trying to forecast the success of the company based on hard, $h^a \in \left\{\underline{h}^a, \bar{h}^a\right\}$, and soft, $h^b \in \left\{\underline{h}^b, \bar{h}^b\right\}$, information. The true probability of h^a being $\bar{h}^a$ or h^b being $\bar{h}^b$ is $.75$ conditional on future success and $.25$ conditional on future failure, where hard and soft signals are conditionally independent. Imagine that the hard and soft signals point in opposite directions, with the hard signal being truly good ($h^a = \bar{h}^a$) and the soft signal being bad ($h^b = \underline{h}^b$). Then, the correct response is to predict the probability of future success to be $1/2$.

People's initial reactions to these signals will vary significantly. However, by Proposition 6, the network of soft-information experts will settle on explaining away the hard information and come to believe the likelihood of future success to be $1/4$. Conversely, the network of hard-information experts will settle on explaining away the soft information and come to believe the likelihood of future success to be $3/4$. The non-experts will settle on explaining away all information and believing the likelihood of future success to be $1/2$. Since some people in the hard- and soft-information networks will start with more moderate (and correct) reactions, in this example social learning intensifies some opinions in the hard- and soft-model networks in addition to hardening them.

With re-labeling, a similar example perhaps sheds light on so-called "epistemic closure" in political debates. Political observers argue that, in recent years, many of beliefs held by conservatives and liberals seem divorced from reality. Pundit Jonathan Chait puts it in the following way:

> the problem is that the [conservative] movement has created its own subculture, and within this subculture, only information from sources controlled by the movement is considered trustworthy or even worth paying attention to.[18]

The key problem, as Chait puts it, is *not* necessarily that liberals are unaware of information provided by conservatives and vice-versa, but rather that they hold shared beliefs that information from the other side of the aisle is not worth grappling with. The analysis in this section shows that this would be a consequence of shared inflexibility in believing information from your own side is trustworthy. Under this interpretation, liberals are aware of conservative information. And they begin with quite diverse opinions on how to interpret conservative information. But, in exchanging interpretations, they end up settling on a shared view that they should not update based on that information.

A final example of networks based on shared models is where the measure $(1 - \delta)$ of the population who initially stick with the default are in one network and the rest of the population are in others. For example, some portion of the population may not devote enough attention to

[18] https://newrepublic.com/article/74492/what-conservative-epistemic-closure-means

an issue to construct their own interpretation of the data beyond the default, nor to exchanging interpretations with others.

When the default is accurate (e.g., in some cases taking scientific consensus at face value), people who adhere to the default end up with more accurate interpretations and beliefs than those in other networks. For example, a 2016 Pew report found that Americans "who care a great deal about GM foods issue expected negative effects from these foods," belying scientific consensus. Similarly, Fernbach et al. (2019) found that people who are extremely opposed to GM foods think they know the most about the safety of those foods, but actually know the least. Such Americans pushed a number of unfounded interpretations of the data, including that eating GM foods caused allergies, cancer, and autism.

5 Managing Networks

Here we ask how someone could try to shape communication networks to her advantage. The network shaper could do this by writing a book, forming groups based on certain shared beliefs/experiences/interests, holding meetings that invite a select group of people, forming social networks, etc. The shaper might also try to prevent certain groups from forming, actively trying to discourage people in one group from speaking to people in another. For example, a manager might insist on being in all meetings with certain subordinates. Or Twitter users might shame those who re-tweet certain arguments.

5.1 Promoting Specific Actions

Suppose first that the network shaper wants to encourage people to take some action in response to the data. For example, in response to a school shooting, the shaper might want to promote gun control, the status quo, or loosening gun restrictions. Formally, consider the case where each person has a finite action space and the shaper's objective is a strictly monotonically increasing function of the fraction of people who choose her ideal action $a^s \in A$. How would the shaper want to structure the network—i.e., the set of models M_i a given person i is exposed to—to maximize this objective?

Proposition 7. *Suppose each person has a finite action space and the network shaper's objective is a strictly monotonically increasing function of the fraction of people who choose her ideal action $a^s \in A$. The network shaper cannot do better than, for every person i, exposing her to all people who would choose a^s in the absence of social learning, and exposing her to nobody else: That is,*

the network-shaper's objective is maximized by setting

$$M_i = \left\{ m \in \bar{M}(h, \mu_0, d, M) : a(\mu(h, m)) = a^s \right\} \tag{1}$$

for all i. The network-shaper's objective continues to be maximized by adding to M_i specified in Eq. (1) *any model m with $\Pr(h|m, \mu_0) < \max_{\tilde{m} \in M_i} \Pr(h|\tilde{m}, \mu_0)$, but it is no longer maximized by adding a model m with $\Pr(h|m, \mu_0) > \max_{\tilde{m} \in M_i} \Pr(h|\tilde{m}, \mu_0)$.*

This result says that the network shaper wants to expose people to all models that support taking action a^s and no other models, except perhaps ones that fit the data plus people's priors worse than the best-fitting model supporting a^s. That is, the shaper wants to form a directed network where everybody listens to people who support action a^s and does not want people to hear good-fitting arguments supporting other actions. If the exact person who would communicate the best-fitting model supporting action a^s were known, they shaper would do as well by having everyone just listen to this person, but realistically the shaper may not be able to identify that person ahead of time. The network-shaper does no worse by exposing everyone to the arguments of people who support action a^s—and it seems plausible to imagine she is able to identify supporters of action a^s.

In a sense, then, this result suggests that a network-shaper who supports a particular action is better off by using a collection of individuals—a platform—to articulate arguments for taking that action than any single individual. A person who wants people to react to recent election results by concluding there is election fraud does better by crowdsourcing arguments from people who have reached this conclusion (and seeing which arguments resonate on social media) than by just leaving it to a single personality to argue. Communities of anti-vaxxers or conspiracy theorists are more persuasive than almost any single person.

We can say more if people are maximally open to persuasion, which we will assume for the rest of this section. Under the optimal network from the perspective of the network shaper, she is only able to get everyone to take her desired action if it is the action people would take in the absence of data—that is, if a^s is the status-quo action $a(\mu_0)$. If a^s is this status-quo action then the best-fitting model among actions that support a^s is the one that says everything that happened is inevitable, $\Pr(h|m, \mu_0) = 1$, which everybody will adopt. If a^s is not this status-quo action, then the best-fitting model that supports a^s has an associated likelihood $\Pr(h|m, \mu_0)$ that is bounded away from one, so a positive measure of individuals will stick with models they came up with in the absence of social learning that have greater associated likelihoods and support sticking with the status quo. The network shaper is at an advantage if she wants everyone to stick with the status quo in response to the data. Importantly, this is true even if the right interpretation of the data is that it supports taking a different action.

To illustrate these results, take the binary example above with $\mu_0(l) = 1/2$ and $h = h^l$. The

network shaper who wants people to choose $a = L$ would want to form networks of small numbers of people who, in the absence of conversation, would choose $a = R$ in a sea of people who would, in the absence of conversation, choose $a = L$. For example, if the networks were of the form "all $\mu(l) \geq .75$ talk to a single person with $\mu(r) > .75$", then that person would end up believing $\mu(l) = .75$. An equally effective network would be one where everybody just listens to anyone who, in the absence of conversation, would choose $a = L$. The most important thing for the network shaper is that people are not exposed to arguments from many people who support the status quo—i.e., those with $\mu(l) \in (.25, .75)$.

More concretely, imagine the status quo is not to take action, the network shaper wants to promote action, and there is an open-to-interpretation event that could lead to action one way or the other. For example, the status quo is some amount of gun control and a school shooting could lead to loosening or tightening gun-control restrictions. Imagine further that the network-shaper supports left-leaning action, e.g., gun control. The people the shaper mosts wants to silence are moderates who argue for inaction, whether or not they are left- or right-leaning. The shaper wants people arguing for left-leaning action to speak and everyone else to listen. And, continuing this logic in a trivial dynamic extension, once all the people arguing for left-leaning action have discussed issues with each other enough to harden beliefs, the shaper is not worried about them having bilateral conversations with reactionaries on the other side—but they would still be wary of them having bilateral converations with those who support the status quo.

5.2 Promoting Shared Models and Actions

We see from the discussion above that, unless the desired action is the status-quo, promoting specific actions typically conflicts with promoting shared models and actions. And a network shaper will sometimes benefit from promoting shared models and actions, for example if she derives benefits from people coordinating on their actions.

To analyze the case where the network shaper wants to promote shared models and actions, suppose the shaper's objective is a strictly monotonically-increasing function of the fraction of people who share what ends up to be the most popular model. The shaper prefers 75% of individuals to hold one model and 25% the other over 60% holding one model and 40% the other, over 50% of individuals holding one model and 50% the other, etc.

Proposition 8. *Suppose the network-shaper's objective is a strictly monotonically-increasing function of the fraction of people who share what ends up to be the most popular model. The network shaper cannot do better than, for every person i, exposing her to all models: That is, the network-*

shaper's objective is maximized by setting

$$M_i = \bar{M}(h, \mu_0, d, M) \tag{2}$$

for all i.

This result says that if the goal is for everyone to end up sharing the same model, the network-shaper cannot do better than encouraging everyone to talk to each other and share their models. When receivers are maximally open to persuasion, this means that the desire for everyone to end up sharing the same model will lead everyone to end up with interpretations that neutralize the data and promote the status-quo action.

6 Applications

6.1 The Evolution and Spread of Misconceptions Through Networks

Why do people believe in misconceptions (e.g., GMOs and vaccines are dangerous) and conspiracy theories (e.g., QAnon) when the Internet and social media also give them access to high-quality information? Echo chambers are a common answer to this question. While people have access to high-quality information, their media diets and social networks only expose them to misinformation and falsehoods. Under this view, falsehoods spread like viruses and crowd out the truth. People hear the same falsehood repeatedly and perhaps then overweight it.

An emerging literature suggests that this misinformation view may be incomplete. Though many people are exposed to lies or falsehoods, they are also exposed to the truth. For instance, Guess et al. (2018) argue that most Americans have diverse media diets, and indeed that social media like Twitter tend to increase the diversity of viewpoints that people are exposed to. Similarly, Bertrand and Kamenica (2020) find that while social attitudes have become stronger predictors of political ideology over time, they have not become stronger predictors of media diet. In addition, Boxell et al. (2017, 2020) find that, while political polarization is increasing, it is not increasing faster for people who extensively use the Internet and social media. Thus, while echo chambers could be a concern, evidence suggests they may not be as widespread a problem as conventional wisdom portrays them to be. Thus, the prevalence of misconceptions in social networks remains a puzzle not fully explained by echo chambers.

Our framework offers a different explanation, highlighting the difference between misconceptions—beliefs that are incorrect due to incorrect interpretations—and misinformation. Within a network, people are exposed to crowdsourced models that evolve to fit the data better and better, which makes them more certain their interpretation of the data is correct and thus more resistant to change.

In sharp contrast to the misinformation and echo-chamber view, bubbles are not about insulating people from certain information; they are about exposing people to interpretations of that data that favor certain beliefs and incoluating them against finding alternative beliefs compelling. Thus, in our framework, the primary impact of bubbles is not to further polarize beliefs, but to harden and make them resistant to change. In particular, even if beliefs react a lot in the immediate aftermath of a big event, bubbles lead members to adopt interpretations that mute and harden their reactions. To stretch the virus analogy, bubbles lead misconceptions to mutate to achieve better fit within a bubble—and people are exposed to a greater degree to mutations within than across bubbles.[19]

While we could illustrate these results by applying the baseline model we presented above, it is more revealing to consider a simple two-period dynamic extension of the analysis under the assumption that everyone is maximally open to persuasion. The key idea the extension highlights is that if networks form endogenously in response to one set of information, those networks will tend to encourage different interpretations of all future information. In other words, endogenous network formation creates strong path dependence in the way people interpret information.

Formally, suppose people begin with the same priors, react to data h_1, and form shared belief networks based on their reactions to h_1. Further suppose that after exchanging models through the network, people's posterior beliefs after interpreting h_1 become their priors in interpreting new data h_2. In interpreting h_2, people share models with others in the shared-belief network that was formed based on common reactions to earlier data h_1. That is, networks are sticky across the two periods: people stay in the shared-belief network that was formed in period 1. For example, people may talk to others who share a similar reaction to well-publicized evidence purporting to show a relationship between vaccines and autism and continue to talk to the same people to make sense of new data that arrives.

The key result from this dynamic extension is that bubbles have lasting consequences on how people interpret subsequent events. By Proposition 3, everyone within a given shared-belief network ends up holding the initial belief closest to the prior within that network in response to data h_1. So everyone within a shared-belief network begins with the same prior entering into the second period where they interpret data h_2. Call this prior belief μ_1^s, which differs across networks s. Since people use the same network to exchange interpretations of h_2, Proposition 1 applies. Social learning maximally mutes and hardens a person's reaction to the data. In other words, everyone

[19]Bowen et al. (2021) provide an alternative model where belief polarization is driven by misperceptions about selective sharing of second-hand information within an echo chamber. In Bowen et al. (2021), disagreement and polarization are driven by different people holding different information (having heterogeneous "information diets" of second-hand information) and not properly accounting for that fact; in our model, disagreement arises even when people share the same information. Their framework sheds light on situations where tons of news is coming out each day and it's hard to keep track of it all (e.g., if there's a war or people are forming beliefs about a new political candidate). We shed light on situations where the basic facts are essentially common knowledge (e.g., the George Floyd murder or the capitol insurrection) and people are primarily exchanging interpretations of those facts.

ends up at the belief they held prior to seeing h_2 with a model that perfectly explains the data: for every person i in shared belief network s, $\mu_i = \mu_1^s$ and $\Pr(h_2|m_i, \mu_1^s) = 1$.

This analysis suggests that networks formed based on shared beliefs may result in beliefs being persistently untethered from data that is open to interpretation. Once misconceptions evolve and harden within a network through crowdsourced interpretations of a high-profile event, members of that network explain subsequent events in a way that makes them consistent with the original interpretation. In other words, a bad take on an event can be very hard to reverse. Importantly, members may originally have disparate (and perhaps realistic) interpretations of subsequent events, but they eventually settle on explanations that neutralize those events.[20]

6.2 When and How to Hold a Meeting

Why do organizations hold so many meetings? Economic models assume meetings are fundamentally about information exchange: One worker holds a piece of information that another does not and exchanging information helps workers adapt to the environment and coordinate their actions (e.g., Dessein and Santos (2006)). Under this account, meetings are essentially no different from other communication technologies (e.g., emails) and are called when workers do not share the same information set. After meetings, workers pull in the same direction, which is better adapted to the full information set.

Organizational scholars view meetings much more broadly. They come in different forms, such as town halls or all hands. They are sometimes about information exchange, but they are also about diagnosing problems, communicating organizational priorities, and exchanging or amplifying views on the right course of action.

This section formalizes such a role for meetings, building on the view put forward in Weick (1995) that sensemaking is a fundamental activity of organizations. In the model, costly meetings are called to help workers make sense of shared information. Meetings allow leaders to control interpretations workers share with each others, and they are called even when workers do not have any new private information. The structure and goal of meetings is not fixed but depends on workers' flow of communication outside meetings and the degree to which the organization prioritizes adaptation versus coordination. In particular, meetings may help workers get on the same page by commonly muting their reaction to data instead of better adapting to the environment.

We consider a similar setting to Dessein and Santos (2006) and Bolton et al. (2013), closely following the latter paper's language and formulation. The environment is parameterized by $\omega \in [0, 1]$, which is not known by the leader or a continuum of followers. Instead, they have a uniform

[20] In the U.S., reactions to the capitol insurrection appear consistent with this pattern: While initial reactions in some corners seemed to break from earlier trends, reactions settled on narratives suggesting there was not much to learn from the event.

prior over ω and interpret data h in terms of what it implies about ω.

The timing of the game, which we flesh out below, is: (1) everyone observes h, (2) the leader announces the organization's strategy $a_L \in [0,1]$ and perhaps holds a meeting to discuss it in light of h, (3) each follower $i \in [0,1]$ chooses an action $a_i \in [0,1]$, and (4) payoffs are realized.

Each follower i has payoff function

$$-\alpha \cdot (a_i - [l_i \cdot a_L + (1 - l_i) \cdot \omega])^2 - \kappa \int_j (a_j - \bar{a})^2 dj,$$

where $\alpha > 0$, $\kappa > 0$, $l_i \in [0,1]$ and $\bar{a} \equiv \int a_j dj$. That is, each follower values (i) taking an action that is aligned with a weighted average of the organization's strategy a_L and the environment and (ii) being part of a well-coordinated organization. To limit the number of cases, we assume that $l_i = 0$ for almost all followers and $l_i = 1$ for positive fraction $\varepsilon \to 0$ of followers.[21] That is, almost all followers care about taking an action that is well-adapted to the state, rather than than taking an action that is aligned with the organization's strategy, and the rest of the followers blindly follow the organization's strategy. By focusing on the case where $l_i = 0$ for fraction $(1 - \varepsilon) \approx 1$ of followers, the analysis below better applies to situations where workers care more about getting things right than about following the leader. The leader's payoff simply aggregates the followers' payoffs:[22]

$$-\alpha \int_i (a_i - [l_i \cdot a_L + (1 - l_i) \cdot \omega])^2 di - \kappa \int_j (a_j - \bar{a})^2 dj.$$

The leader and followers share the same default model. While the leader is dogmatic the default model is correct, followers may move away from it by sensemaking on their own and with fellow followers.

Because Ω in this example is the full unit interval, we for simplicity limit the set of models M followers could consider to be finite. We assume M always includes (i) the default model d, (ii) the best-fitting model m^{bf} that induces the same beliefs as d (i.e., $\mu(h, m^{bf}) = \mu(h, d)$), (iii) a model that says the history is inevitable in hindsight (i.e., a model m such that $\Pr(h|m, \mu_0) = 1$), and (iv) at least one model m satisfying $\Pr(h|m, \mu_0) \in (\Pr(h|m, d), \Pr(h|m^{bf}, \mu_0))$ and $\mu(h, m) \neq$

[21] Having some followers blindly follow the organization's strategy induces a cost to the leader of announcing a different strategy from what she thinks is subjectively optimal. There are other ways to generate such a cost, e.g., by assuming as Bolton et al. (2013) do that followers and leaders value being part of an organization that is well-adapted to its environment. We take the approach we do because it is analytically simpler for our purposes than such other approaches, but our qualitative results do not hinge on our precise formulation.

[22] For simplicity, we assume the leader evaluates her expected payoff according to her own expectation and not followers' subjective expectations. For example, the leader has an incentive for followers' actions to be well-adapted to the leader's view of the environment, but does not directly care whether the followers believe their actions are well-adapted to the environment. Introducing the latter force could provide an additional reason why leaders want to hold meetings in our framework: to get followers on board with the direction of the organization, even when getting followers on board does not influence their actions.

$\mu(h, d)$. We also for simplicity assume that m^{bf} fits better than all models in M except for the model that says the history is inevitable in hindsight.

If the leader does not hold a meeting, then workers make sense of h in their own networks. Holding a meeting costs the leader a positive amount c that is vanishingly small. By holding a meeting, the leader is able to perfectly control the set of models each worker is exposed to, M_i, by influencing the flow of communication between followers.

Proposition 9. *In the leader-follower example of this section:*

1. *If information h is closed to interpretation or followers always stick with their default interpretation of the information absent persuasion ($\delta = 1$), then the leader never holds a meeting. In this case, $a_L = \mathbb{E}_{\mu(h,d)}[\omega]$ for all h, and $a_i = a_L$ for all i.*

2. *Otherwise, the leader may hold a meeting.*

   (a) *If the weight placed on coordination (κ) is sufficiently large or if h is uninformative under the default model in the sense that $\mathbb{E}_{\mu(h,d)}[\omega] = \mathbb{E}_{\mu_0}[\omega] \equiv \omega_0$, then the leader calls a meeting whenever some followers take an action other than ω_0 absent a meeting, for example because followers are in a network where not everyone talks to everyone else. Additionally, in this case (i) an optimal meeting features open communication ($M_i = M$ for all i), (ii) $a_L = \omega_0$, and (iii) $a_i = \omega_0$ for all i.*

   (b) *If the weight placed on adaptation (α) is sufficiently large and followers should react to the information under the default model in the sense that $\mathbb{E}_{\mu(h,d)}[\omega] \neq \mathbb{E}_{\mu_0}[\omega]$, then the leader calls a meeting whenever too many followers take an action other than $\mathbb{E}_{\mu(h,d)}[\omega]$ absent a meeting, for example, because they talk to others who supply a model that says the history is inevitable in hindsight. Additionally, in this case (i) an optimal meeting features directed communication with $M_i \neq M$, (ii) $a_L \neq \omega_0$, and (iii) not all followers take the same action.*

The first part of Proposition 9 says that, when data is closed to interpretation or followers do not try to make sense of the data on their own, then there is no need for the leader to call a meeting to discuss the organization's strategic response to publicly available data. The leader just announces her strategic response, which varies one-for-one with the leader's reaction to the data.

The second part of the proposition shows how the leader's reaction is very different when data is open to interpretation and followers try to make sense of it on their own. Meetings then allow leaders to better control interpretations followers share with each other. If, in the leader's mind, followers are reacting to data when they should not be or if the organization places a sufficiently large weight on coordination, then the leader calls a meeting which features open communication:

everyone shares their view of what the event means for the organization. While opinions will be voiced that leaders do not agree with, at the end of the day everyone will share a view that the event teaches them little that they did not already know. Thus, the status quo will prevail. In this case, the leader's strategic response to publicly available data may be muted relative to her private response: if she believes that she cannot persuade enough followers in her desired course of action through a meeting, her next-best alternative is to ensure coordination by structuring the meeting to neutralize the data. This may be one reason why informal (e.g., relational) contracts are "hard to build *and change*" (emphasis added, Gibbons and Henderson (2012b)).

On the other hand, if too many followers are underreacting to the data or the organization places a sufficiently large weight on adaptation, then the leader calls a meeting which features a sort of *persuasive campaign* where leaders ensure that the loudest voices are those that interpret the event in a way consistent with its view of the optimal course of action $\mathbb{E}_{\mu(h,d)}[\omega]$. In this case, leadership would worry about certain interpretations being more compelling than the desired one. While not everyone ends up on board with the shift in strategy from the status quo of ω_0, as many will be on board as possible. Per Proposition 5 there is also desire to hold the meeting as soon as possible, before workers can share interpretations with each other on their own. Indeed, the proof of Proposition 9 establishes that the benefits of holding a meeting that prioritizes adaptation (through a persuasive campaign) versus coordination (through open communication) is decreasing in the fraction of followers who adopt the "everything is obvious in hindsight" model prior to the meeting; the longer the delay before holding the meeting, the greater the risk of this model spreading among followers.

7 Discussion

This paper is just a first step to studying the social transmission of models. While we assume people costlessly exchange models with others, in many cases people devote effort, attention, and time to exposing themselves to new models for reasons of curiosity, identity, and instrumentality. How does incorporating a realistic demand function for models influence, for example, the way networks are structured? Appendix B presents one extension along these lines where a platform designer promotes engagement by exposing individuals to models that provide good explanations and/or to explanations that justify their pre-existing beliefs. The appendix shows that such platforms are not a force towards truth but of hardening people's views.

The framework also admits further applications. For example, if a manager wants to organize teams to help her arrive at a realistic interpretation of the data, how would she do it? Would she like to construct teams of advocates to particular positions (i.e., shared-action networks)? Of teams who tend to reach similar conclusions (i.e., shared-belief networks)? Of teams who look at the data

in similar ways (i.e., shared-model networks)? A loose intuition that arises from the framework which we have yet to formalize is that, reminiscent of Hong and Page (2001) and per the results in Section 4, a good manager is able to harness systematically diverse viewpoints on how to interpret data to reach more accurate conclusions than if she only heard a single viewpoint. On the other hand, per the results in Section 3, even a good manager does not benefit from hearing viewpoints that differ not because of systematically different ways of looking at the data but rather a tendency to reach systematically different *conclusions* from the data. In the Venture Capital context, it's ok to have people who focus on the idea and people who focus on the team but not people who want to invest and people who want to pass.

A Proofs

Proof of Proposition 1. That social learning hardens every person's reaction to the data is immediate from how models are selected. That social learning leads everyone to end up at their prior follows from the fact that someone will come up with and communicate the model m that $\pi_m(h|\omega) = 1$ for all $\omega \in \Omega$, which will beat all other models (since $\Pr(h|m, \mu_0) = 1$) and leads to $\mu(h, m) = \mu_0$.

□

Proof of Proposition 2.

1. That expanding person i's network hardens her reaction to data follows from the simple fact that $\max_{m \in M^e} \Pr(h|m, \mu_0) \geq \max_{m \in M} \Pr(h|m, \mu_0)$ whenever $M^e \supset M$.

2. For every $\tilde{\mu}$, there exists a model $m(\tilde{\mu})$ supporting that belief that is less compelling than the model m_i the person would adopt prior to the network expansion: for example, take the model
$$\pi_{m(\tilde{\mu})}(h|\omega) = \frac{\tilde{\mu}(\omega)}{\mu_0(\omega)} \cdot (\Pr(h|m_i, \mu_0) - \varepsilon)$$
for all $\omega \in \Omega$ and for $\varepsilon > 0$ small.

3. This follows from the fact that there are always multiple (in fact infinite) models that induce a given belief $\tilde{\mu}$, and they do not all have the same fit.

□

Proof of Proposition 3. Someone who holds the most moderate initial belief in a network will come up with and communicate the best-fitting model that leads to that belief—that is, the model m leading to that belief that maximizes $\Pr(h|\cdot, \mu_0)$. By Lemma 1, this model will beat out all others in the network.

□

Proof of Proposition 4.

1. That expanding person i's network hardens her reaction to data follows from the simple fact that $\max_{m \in M^e} \Pr(h|m, \mu_0) \geq \max_{m \in M} \Pr(h|m, \mu_0)$ whenever $M^e \supset M$. That expanding person i's network if anything mutes her reaction to the data follows from the fact that m_i is the best-fitting model inducing μ_i, which fits better than any model inducing a less-temperate belief (by Lemma 1).

2. This part is a corollary of Proposition 3.

3. If $\tilde{\mu}$ is less temperate than μ_i, then it cannot impact person i's final beliefs because μ_i is supported by the best-fitting model inducing that belief (by Proposition 3), which provides a better fit than any model supporting any less temperate belief (by Lemma 1).

4. The "only if" part follows from the proof of part (3). The "if" part follows from the fact that, when $\tilde{\mu}$ is more temperate than μ_i, the best-fitting model supporting $\tilde{\mu}$ fits better than the best-fitting model m_i supporting μ_i (by Lemma 1).

□

Proof of Proposition 5. Weak exposure to belief $\tilde{\mu}$ prior to social learning impacts the person's final beliefs if and only if the person finds $m(\tilde{\mu})$ more compelling than the model m_i' she currently has in mind supporting belief μ_i': that is, if and only if

$$\Pr(h|m(\tilde{\mu}), \mu_0) > \Pr(h|m_i', \mu_0). \tag{3}$$

Weak exposure to belief $\tilde{\mu}$ through network expansion impacts the person's final beliefs if and only if the person finds $m(\tilde{\mu})$ more compelling than the best-fitting model among those represented in shared-belief network $s(\mu_i')$: that is, if and only if

$$\Pr(h|m(\tilde{\mu}), \mu_0) > \max_{m' \in \bigcup_{\mu \in s(\mu_i')} M(\mu)} \Pr(h|m', \mu_0). \tag{4}$$

The result follows from the right-hand-side of inequality (4) being larger than the right-hand-side of inequality (3).

A similar proof applies to the case of strong exposure to beliefs, replacing the left-hand-sides of inequalities (3) and (4) with $\max_{m' \in M(\tilde{\mu})} \Pr(h|m', \mu_0)$. □

Proof of Proposition 6. Recall that

$$c^A = \left\{ m \in \bar{M}(h, \mu_0, d, M) : \pi_m(h^a, h^b|\omega) = \pi_m(h^b|\omega) \cdot \pi_{m^{fa}}(h^a|\omega) \; \forall \; \omega \in \Omega \right\}.$$

Clearly, the best fitting model in c^A is $\pi_m(h^a, h^b|\omega) = 1 \cdot \pi_{m^{fa}}(h^a|\omega) = \pi_{m^{fa}}(h^a|\omega)$ for all $\omega \in \Omega$. Similarly, the best fitting model in c^B is $\pi_m(h^a, h^b|\omega) = 1 \cdot \pi_{m^{fb}}(h^b|\omega) = \pi_{m^{fb}}(h^b|\omega)$ for all $\omega \in \Omega$. Finally, the best fitting model in c^O is $\pi_m(h^a, h^b|\omega) = 1$ for all $\omega \in \Omega$. By assumption, someone in each network will propose the associated best-fitting models which all network members will end up adopting. The final beliefs μ_i follow.

□

Proof of Proposition 7. The network-shaper's objective is clearly maximized by exposing everybody to the best-fitting model that supports action a^s and exposing them to no other models. The network-shaper does no worse by exposing people to all models specified in Eq. (1) (i.e., all models that support action a^s), since this includes the best-fitting one and no models that support other actions. That is, everybody's behavior is the same whether they are only exposed to the best-fitting

model that supports a^s or models specified in Eq. (1). This remains true if we add to models specified in (1) any model m with $\Pr(h|m,\mu_0) < \max_{\tilde{m}\in M_i} \Pr(h|\tilde{m},\mu_0)$, since nobody will adopt such a model. However, the network-shaper's payoff is strictly worse if we add to models specified in (1) any model m with $\Pr(h|m,\mu_0) > \max_{\tilde{m}\in M_i} \Pr(h|\tilde{m},\mu_0)$, since anybody who would've adopted a model in M_i will instead adopt this model which supports taking an action other than a^s.

□

Proof of Proposition 8. If everybody is exposed to $\bar{M}(h,\mu_0,d,M)$, then everybody will also end up adopting the model in that set that maximizes $\Pr(h|\cdot,\mu_0)$. The network-shaper cannot do better, since everyone will end up sharing the same model.

□

Proof of Proposition 9. For the first case, it's obvious that the leader never holds a meeting because holding a meeting costs $c > 0$ and does not influence beliefs and decisions when information is closed to interpretation or when followers always stick with their default interpretation of the information absent persuasion. Since $a_L = \mathbb{E}_{\mu(h,d)}[\omega]$ implies $a_i = a_L$ for all i (this is obvious for followers who blindly follow a_L and other followers set $a_i = l \cdot a_L + (1-l) \cdot \mathbb{E}_{\mu(h,d)}[\omega] = a_L$), it remains to show in this case that $a_L = \mathbb{E}_{\mu(h,d)}[\omega]$. Setting $a_L = \mathbb{E}_{\mu(h,d)}[\omega]$ uniquely maximizes the coordination term, $-\int_j (a_j - \bar{a})dj$, of the leader's payoff since everyone coordinates on a_L. Since simple algebra shows that a_L doesn't influence the adaptation term, $\int_i -(a_i - [l_i \cdot a_L + (1-l_i)\cdot\omega])^2 di$, it is optimal for the leader to set $a_L = \mathbb{E}_{\mu(h,d)}[\omega]$.

For the first part of the second case, optimizing the leader's payoff becomes equivalent to maximizing the coordination term, $\int_j (a_j - \bar{a})^2 dj$, when the weight placed on coordination κ is sufficiently large. Given that a positive fraction of followers initially adopt the perfectly-fitting neutralizing model, the only way for all followers to perfectly coordinate their actions is for them all to take $a_i = \omega_0$. This is implemented by followers being exposed to all models, either with open communication absent a meeting or with open communication in a meeting. This is also optimal from the point of view of the leader when h is uninformative under the default model in the sense that $\mathbb{E}_{\mu(h,d)}[\omega] = \omega_0$. The leader does better by holding a meeting than not whenever some followers would adopt a model that implies a belief other than μ_0 absent a meeting.

For the last part, if followers are exposed to all models ($M_i = M$ for all i), then they perfectly coordinate their actions and the leader's payoff approximately equals

$$-\alpha \mathbb{E}_{\mu(h,d)} \int_i (a_i - \omega)^2 di = -\alpha \mathbb{E}_{\mu(h,d)} \int_i (\omega_0 - \omega)^2 di \tag{5}$$

since $l_i = 0$ for almost all followers. If followers are instead all exposed to only models supporting

$a_i = \mathbb{E}_{\mu(h,d)}[\omega]$ (i.e., $M_i = \left\{d, m^{bf}\right\}$ for all i), then the leader's payoff approximately equals

$$-\alpha\left[\mathbb{E}_{\mu(h,d)}\rho\int_i(\mathbb{E}_{\mu(h,d)}[\omega]-\omega)^2di+(1-\rho)\int_i(\omega_0-\omega)^2di\right]-\kappa\int_j(a_j-\rho\mathbb{E}_{\mu(h,d)}[\omega]-(1-\rho)\omega_0)^2dj, \tag{6}$$

where ρ equals the fraction of followers who are persuadable by m^{bf} (i.e., fraction $1-\rho$ are the fraction with the initial reaction to adopt the perfectly-fitting neutralizing model). Since the first term of (6) is larger than (5) when $\mathbb{E}_{\mu(h,d)}[\omega] \neq \omega_0$, in this case the leader holds a meeting that features directed communication whenever α is sufficiently large. Such a meeting will clearly be better than not holding a meeting whenever followers whose initial reaction to the data differs from $\mu(h,d)$ are not exposed to m^{bf} absent a meeting or are exposed to the model that says the history is inevitable in hindsight.[23]

□

B Promoting Engagement on a Platform

Suppose the network-shaper is a platform designer who wants to encourage engagement on the platform. To incorporate engagement in a simple way, we extend the model in the spirit of Mullainathan and Shleifer (2005) to suppose people don't like having their beliefs disconfirmed: The engagement of any given person i is decreasing in the distance between pre-social-learning beliefs μ_i' and post-social-learning beliefs μ_i. At the same time, we suppose people like learning new arguments that support their positions: The engagement of any given person i is increasing in the degree to which the person's beliefs are hardened following social learning—that is, in the distance between how well the person's adopted model following social learning explains the data, $\Pr(h|m_i,\mu_0)$, and how well her adopted model prior to social learning explains the data, $\Pr(h|m_i',\mu_0)$.

Overall, suppose the engagement of person i, e_i, is a weighted average of these two factors:

$$e_i = \beta\cdot[-d(\mu_i,\mu_i')]+(1-\beta)\cdot\left(\Pr(h|m_i,\mu_0)-\Pr(h|m_i',\mu_0)\right), \tag{7}$$

[23] To see when else the leader wants to hold such a meeting, (6) minus (5) equals:

$$-\alpha\rho\mathbb{E}_{\mu(h,d)}\left[(\mathbb{E}_{\mu(h,d)}[\omega]-\omega)^2-(\omega_0-\omega)^2\right]-\kappa\left[\int_j(a_j-\rho\mathbb{E}_{\mu(h,d)}[\omega]-(1-\rho)\omega_0)^2dj\right],$$

which, after some algebra, equals $\alpha\rho(\mathbb{E}_{\mu(h,d)}[\omega]-\omega_0)^2-\kappa\rho(1-\rho)(\mathbb{E}_{\mu(h,d)}[\omega]-\omega_0)^2$. So a meeting featuring directed communication is optimal whenever $\alpha>\kappa(1-\rho)$. This reveals that a leader is more likely to call a meeting to encourage followers to take an action different from ω_0 the greater the fraction of followers who are persuadable to take such an action—that is, the smaller the fraction of followers who, prior to the meeting, are hardened in their views that the data tells them little they didn't already know.

where $\beta \in (0,1)$ and $d(\cdot)$ is some distance metric between beliefs. We assume that for every person i, the platform designer is able to observe the person's pre-social-learning beliefs μ_i' and then select the set of models the person is exposed to among those others have come up with, M_i, to maximize e_i.

While the general problem of maximizing e_i appears intractable without further assumptions, it's easy to solve for what happens in the limit cases, which also gives a flavor for the general solution. If engagement mostly depends on a person's beliefs not being disconfirmed, $\beta \approx 1$, then $M_i = \left\{m \in \bar{M}(h, \mu_0, d, M) : \mu(h, m) = \mu_i'\right\}$: that is, the designer will implement an extreme form of shared belief networks where a person is exposed only to models that confirm her pre-social-learning beliefs. In this case, which might hold when people's identity is connected to a specific belief (e.g., on which political candidate would govern better), a platform takes a person's revealed belief and returns other arguments supporting that belief. At the end of the day, the platform leads a person to more strongly hold any belief she started with.

If, on the other hand, engagement mostly depends on the person being better able to explain data following social learning, $\beta \approx 0$, then $M_i = \bar{M}(h, \mu_0, d, M)$. In this case, which might hold when a person is curious about an issue but doesn't identify strongly with a particular position, the person is exposed to everyone else's model. At the end of the day, the platform leads a person to react less to the data by exposing them to models where the data feels obvious in hindsight.

In both of the extremes, the platform is not a force towards truth but of hardening people's positions. And in the latter extreme, the platform leads people's beliefs to converge but to become completely untethered to the data.

References

Aina, Chiara, "Tailored Stories," 2021.

Akbarpour, Mohammad, Suraj Malladi, and Amin Saberi, "Just a Few Seeds More: Value of Network Information for Diffusion," 2020.

Alesina, Alberto, Armando Miano, and Stefanie Stantcheva, "The Polarization of Reality," in "AEA Papers and Proceedings," Vol. 110 2020, pp. 324–28.

Banerjee, Abhijit V, "A Simple Model of Herd Behavior," *The Quarterly Journal of Economics*, 1992, *107* (3), 797–817.

Bénabou, Roland, Armin Falk, and Jean Tirole, "Narratives, Imperatives, and Moral Reasoning," Technical Report, National Bureau of Economic Research 2018.

Bertrand, Marianne and Emir Kamenica, "Coming Apart? Cultural Distances in the United States Over Time," 2020.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch, "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades," *Journal of Political Economy*, 1992, *100* (5), 992–1026.

Bolton, Patrick, Markus K Brunnermeier, and Laura Veldkamp, "Leadership, Coordination, and Corporate Culture," *Review of Economic Studies*, 2013, *80* (2), 512–537.

Bowen, Renee, Danil Dmitriev, and Simone Galperti, "Learning from Shared News: When Abundant Information Leads to Belief Polarization," 2021.

Boxell, Levi, Matthew Gentzkow, and Jesse M Shapiro, "Greater Internet Use is Not Associated with Faster Growth in Political Polarization Among US Demographic Groups," *Proceedings of the National Academy of Sciences*, 2017, *114* (40), 10612–10617.

_ , _ , and _ , "Cross-Country Trends in Affective Polarization," Technical Report, National Bureau of Economic Research 2020.

Burnstein, Eugene and Amiram Vinokur, "Persuasive Argumentation and Social Comparison as Determinants of Attitude Polarization," *Journal of Experimental Social Psychology*, 1977, *13* (4), 315–332.

Chater, Nick and George Loewenstein, "The Under-Appreciated Drive for Sense-Making," *Journal of Economic Behavior & Organization*, 2016, *126*, 137–154.

DeGroot, Morris H, "Reaching a Consensus," *Journal of the American Statistical Association*, 1974, *69* (345), 118–121.

DeMarzo, Peter M, Dimitri Vayanos, and Jeffrey Zwiebel, "Persuasion Bias, Social Influence, and Unidimensional Opinions," *The Quarterly Journal of Economics*, 2003, *118* (3), 909–968.

Dessein, Wouter and Tano Santos, "Adaptive organizations," *Journal of Political Economy*, 2006, *114* (5), 956–995.

Eliaz, Kfir and Ran Spiegler, "A Model of Competing Narratives," *American Economic Review*, 2020, *110* (12), 3786–3816.

Enke, Benjamin and Florian Zimmermann, "Correlation Neglect in Belief Formation," *The Review of Economic Studies*, 2019, *86* (1), 313–332.

Eyster, Erik and Matthew Rabin, "Naive Herding in Rich-Information Settings," *American Economic Journal: Microeconomics*, 2010, *2* (4), 221–43.

_ **and** _ , "Extensive Imitation is Irrational and Harmful," *The Quarterly Journal of Economics*, 2014, *129* (4), 1861–1898.

Fernbach, Philip M, Nicholas Light, Sydney E Scott, Yoel Inbar, and Paul Rozin, "Extreme Opponents of Genetically Modified Foods Know the Least but Think They Know the Most," *Nature Human Behaviour*, 2019, *3* (3), 251–256.

Froeb, Luke M, Bernhard Ganglmair, and Steven Tschantz, "Adversarial Decision Making: Choosing Between Models Constructed by Interested Parties," *The Journal of Law and Economics*, 2016, *59* (3), 527–548.

Gagnon-Bartsch, Tristan and Matthew Rabin, "Naive social learning, mislearning, and unlearning," 2016.

Gentzkow, Matthew and Jesse M Shapiro, "Ideological Segregation Online and Offline," *The Quarterly Journal of Economics*, 2011, *126* (4), 1799–1839.

Gershman, Samuel J, "How to Never be Wrong," *Psychonomic Bulletin & Review*, 2019, *26* (1), 13–28.

Gibbons, Robert, "Deals That Start When You Sign Them," 2021.

_ **and Laurence Prusak**, "Knowledge, Stories, and Culture in Organizations," in "AEA Papers and Proceedings," Vol. 110 2020, pp. 187–92.

_ **and Rebecca Henderson**, "Relational Contracts and Organizational Capabilities," *Organization science*, 2012, *23* (5), 1350–1364.

_ **and** _ , "What Do Managers Do? Exploring Persistent Performance Differences among Seemingly Similar Enterprises," 2012.

Golub, Benjamin and Evan Sadler, "Learning in Social Networks," in "The Oxford Handbook of the Economics of Networks" 2016.

_ **and Matthew O Jackson**, "Naive Learning in Social Networks and the Wisdom of Crowds," *American Economic Journal: Microeconomics*, 2010, *2* (1), 112–49.

Guess, Andrew, Brendan Nyhan, Benjamin Lyons, and Jason Reifler, "Avoiding the Echo Chamber About Echo Chambers," *Knight Foundation*, 2018, *2*.

Hirshleifer, David, "Presidential Address: Social Transmission Bias in Economics and Finance," *The Journal of Finance*, 2020, *75* (4), 1779–1831.

Hong, Lu and Scott E Page, "Problem Solving by Heterogeneous Agents," *Journal of Economic Theory*, 2001, *97* (1), 123–163.

Lin, Yu-Ru and Wen-Ting Chung, "The Dynamics of Twitter Users' Gun Narratives Across Major Mass Shooting Events," *Humanities and Social Sciences Communications*, 2020, *7* (1), 1–16.

Mullainathan, Sendhil and Andrei Shleifer, "The Market for News," *American Economic Review*, 2005, *95* (4), 1031–1053.

_ , Joshua Schwartzstein, and Andrei Shleifer, "Coarse Thinking and Persuasion," *The Quarterly journal of economics*, 2008, *123* (2), 577–619.

Murphy, Kevin M and Andrei Shleifer, "Persuasion in politics," *American Economic Review*, 2004, *94* (2), 435–439.

Quattrone, George A and Edward E Jones, "The Perception of Variability Within In-Groups and Out-Groups: Implications for the Law of Small Numbers.," *Journal of personality and social psychology*, 1980, *38* (1), 141.

Roux, Nicolas and Joel Sobel, "Group Polarization in a Model of Information Aggregation," *American Economic Journal: Microeconomics*, 2015, *7* (4), 202–32.

Schkade, David, Cass R Sunstein, and Daniel Kahneman, "Deliberating About Dollars: The Severity Shift," *Colum. L. Rev.*, 2000, *100*, 1139.

_ , _ , and Reid Hastie, "What Happened on Deliberation Day," *Calif. L. Rev.*, 2007, *95*, 915.

Schulz, Laura E and Jessica Sommerville, "God Does not Play Dice: Causal Determinism and Preschoolers' Causal Inferences," *Child Development*, 2006, *77* (2), 427–442.

Schwartzstein, Joshua and Adi Sunderam, "Using Models to Persuade," *American Economic Review*, 2021, *111* (1), 276–323.

Shiller, Robert J, "Narrative Economics," *American Economic Review*, 2017, *107* (4), 967–1004.

_ , *Narrative Economics: How Stories go Viral and Drive Major Economic Events*, Princeton University Press, 2020.

Smith, Lones and Peter Sørensen, "Pathological Outcomes of Observational Learning," *Econometrica*, 2000, *68* (2), 371–398.

Weick, Karl E, *Sensemaking in Organizations*, Vol. 3, Sage, 1995.