New England Journal of Public Policy

Volume 34 | Issue 2 Article 4

10-19-2022

Bounded Confidence: How AI Could Exacerbate Social Media's Homophily Problem

Dylan Weber
Changing Character of War Centre, Pembroke College, Oxford University and Artis Research

Scott Atran
Changing Character of War Centre, Pembroke College, Oxford University and Artis Research

Rich Davis
Changing Character of War Centre, Pembroke College, Oxford University and Artis Research

Bounded Confidence: How AI Could Exacerbate Social Media's Homophily Problem

Dylan Weber, Scott Atran, and Rich Davis
Changing Character of War Centre, Pembroke College, Oxford University and Artis Research

Abstract

The advent of the Internet was heralded as a revolutionary development in the democratization of information. It has emerged, however, that online discourse on social media tends to narrow the information landscape of its users. This dynamic is driven by the propensity of the network structure of social media to tend toward *homophily;* users strongly prefer to interact with content and other users that are similar to them. We review the considerable evidence for the ubiquity of homophily in social media, discuss some possible mechanisms for this phenomenon, and present some observed and hypothesized effects. We also discuss how the homophilic structure of social media makes it uniquely vulnerable to artificial-intelligence-driven, automated influence campaigns.

Rich Davis and Scott Atran are the founding fellows of Artis Research, where Dr. Davis serves as Chief Executive Officer and Professor Atran serves as Director of Research. Both are Research Fellows at the Changing Character of War Centre. Dylan Weber is a Research Fellow at Artis Research, where he additionally serves as Director of Artificial Intelligence. Dr. Weber is also a Research Fellow at the Changing Character of War Centre.

Since the advent of the Internet, the literature has debated how such a large shift in the information landscape will affect the function of democracies. In the early years of the twenty-first century, many argued that since the Internet offered the possibility of larger, more direct discussion among the populace as well as greatly enhanced access to information, its adoption would be a boon for democratic society.[1] The prevailing sentiment at the time (as it is now) was that a functioning democracy is synonymous with healthy deliberation. Others cautioned, however, that an information sphere lacking intermediaries and offering a virtually infinite array of information sources would foster deliberation that is anything but healthy.[2] These scholars cautioned that given the array of choice that the Internet offers, finite time to spend consuming information, and the well-established psychological tendencies of confirmation bias and selective exposure, the Internet would not broaden the information horizons of its users but narrow them. They predicted that given the choice, people would spend their time on the Internet interacting with those with whom they already agreed and consuming information from sources that confirmed their prior views; in other words, they predicted that the interaction structure of the Internet would tend toward *homophily*. In the past fifteen years, the advent of social media and its associated meta data has opened the door to empirical answers to these questions and the answers are stark.

In the first section we discuss the considerable evidence that the network structure of social media does tend toward homophily, present some possible mechanisms for that tendency, and outline some observed and hypothesized effects. In the second section we employ mathematical modeling to illustrate how networks that are prone to homophily are also susceptible to polarizing agents. Finally, we'll discuss how artificial intelligence techniques could be leveraged to deploy automated polarizing agents at scale to exploit the homophilic structure of social media.

Social Media's Homophily Problem

Homophilic Structure of Social Media

Since its advent, there has been concern in academic circles about the formation of "echo chambers" on social media.[3] But until the past seven years or so, there has been a dearth of empirical investigation into the question whether such communities were actually emerging. These concerns were catapulted into the mainstream during the 2016 US presidential election, prompting a large number of studies looking into the question by leveraging large social media datasets. Because of the variation in specific methods and in the definition of "echo chamber" across these studies, we first offer some definitions to put all the findings on common ground.

Definition 1 *A network is a collection of nodes and a collection of edges that connect the nodes.*

Definition 2 *Given a network and a quantity defined on its nodes that is a priori independent of the network edge structure, we say that the network is homophilic with respect to the quantity if nodes are more likely to be connected in the network if they have a similar value of the quantity.*

In short, a network is homophilic if "users of a feather connect together." All the research we review into identifying the existence of echo chambers on social media is unified in the sense that it aims to study homophily in the networks with respect to some measure of ideology on a given issue or issues (usually political). These studies also are generally unified in their methods. First, they collect a large social media dataset on an issue (or issues) in question. Next, they define a methodology for quantifying each user's ideology on the issue in question. Finally, they define a scheme for structuring their dataset into a network and quantifying homophily with respect to the ideology measure—this step is usually accomplished through some type of clustering of the

network interactions in the ideology space of the users. The network structure is always drawn from the inherent network structure present in social media data from the various interactions users can have with each other (e.g., following, friendship, commenting, retweeting). The ideology measure is usually computed using one of two general strategies. The first, which we call the "link to labels" strategy, involves identifying a list of social media entities with wide followings (e.g., politicians, news outlets, Facebook pages) and asking a group of subject-matter experts to label the ideology values of each entity. The ideology of a given user in the dataset can then be computed from the ideology values of the labeled entities to which the user is connected in the network structure. The second, which we call the "scaling" strategy, uses features independent of the network structure that is being examined and a traditional statistical or deep learning model to fit an ideology value for each user. Surprisingly, even given the range of issues that they examine, this body of research is also remarkably unified in its findings; social media has a strikingly homophilic structure.

One of the first quantitative studies to examine homophily in social media data using a scaling strategy was conducted by a team led by Pablo Barberá.[4] They introduce a statistical model that uses the user-follows-politician bipartite network on Facebook as the main feature to fit ideology values for the users. Using this strategy, they found homophily in the friendship network drawn from a Facebook dataset that concerns the 2012 presidential election, which demonstrates that this tendency was not unique to the 2016 race. Using the same methodology, Barberá and colleagues find a similar trend across a collection of other political issues, including the 2013 government shutdown, minimum wage, and marriage equality.[5] In two studies of Twitter, Kiran Garimella and colleagues study Twitter and define the ideology of a user through the link to labels strategy in the first study, and the scaling method introduced by Barberá's team in the second.[6] In addition to users, they quantify the ideology of *tweets* produced by users that include a link to news outlets through the link to labels strategy. They find pronounced homophily in Twitter networks discussing gun control, Obamacare, and abortion; users strongly tended to both produce and consume tweets whose ideology matched with their own as well as interact with users whose ideology measure was close to their own. Interestingly, when this analysis was repeated in networks corresponding to several nonpolitical topics, homophily was not observed; this effect was also observed by Barberá's team.[7]

Cinelli and colleagues also find pronounced homophily in the Twitter conversation about abortion as well as vaccines using the link to labels strategy.[8] In both conversations they find two well-defined groups of users with disparate ideology values who strongly preferred to interact only within their groups. They extend this analysis to Facebook, with similar results. But when the same methodology was repeated using data from Reddit and Gab, users tended to cluster into only one group. On Reddit this group reflected a moderately liberal ideology, on Gab the singular group was strongly conservative. This result suggests that entire social platforms are echo chambers with respect to certain issues. The existence of homophily in the vaccination debate on Twitter is confirmed by Mønsted and colleagues using a scaling strategy, and the analysis is extended to Facebook by Schmidt and colleagues using the link to labels strategy where homophily is identified as well.[9]

Homophily is not limited to a specific set of topics. For example, the authors of "Exposure to Ideologically Diverse News and Opinion on Facebook," "Quantifying Social Media's Political Space," and "Sharing Political News: The Balancing Act of Intimacy and Socialization in Selective Exposure" do not limit their datasets to content pertaining to a specific issue and instead attempt to capture the political space on Facebook writ large.[10] Even in this much-less-focused domain, all three studies find pronounced homophily. It does not appear, however, that the tendency toward

homophily is constricted to topics that are explicitly political. In "Echo Chambers on Facebook," "Debunking in a World of Tribes," and "Homophily and Polarization in the Age of Misinformation," the authors examine a large Facebook dataset consisting of all the content from a large list of Facebook pages labeled by subject matter experts as "science" or "conspiracy."[11] They define the ideology of a user through the proportion of their activity directed at science or conspiracy pages and define a user as "polarized" if at least 95 percent of their interactions are with pages of one type. They find that of users who interacted with a conspiracy page that 91.53 percent were polarized toward conspiracy and of users who interacted with a science page that 76.79 percent were polarized toward science. Additionally, they find homophily in the Facebook friendship network with respect to this ideology measure— those polarized to conspiracy were very unlikely to be friends with those polarized toward science and vice versa. A similar trend is found in analyses of the news consumption patterns of users on Facebook: users tend to interact only with a small group of similarly aligned outlets in lieu of all other news sources. Users who consume the same group of pages are much more likely to interact with each other.[12] Finally, the tendency toward homophily on social media has been observed in multiple locales outside the United States as demonstrated by Grömping, Cota, and colleagues, and Barberá and colleagues.[13]

Mechanisms of Homophily

There is much debate about what could be driving the striking tendency toward homophily observed in social media, and research into such mechanisms is still in its early stages. Some point to the algorithmic curation of content as a main culprit; they caution that it could cause users to be exposed only to content for which they have previously demonstrated an affinity, otherwise known as a "filter bubble."[14] But there is a growing body of evidence that this effect, though it may exist, might be less pronounced than feared.[15] Garret offers a good review.[16]

Bakshy and colleagues find that "there is on average slightly less cross-cutting content: conservatives see approximately 5 percent less cross-cutting content compared to what friends share, while liberals see about 8 percent less ideologically diverse content." In the same study, the authors find that while individuals may be exposed to cross-cutting content, they engage with it at much lower rates.[17] Garret admits strong evidence for this phenomena as well, coining the term "engagement echo chamber." So while algorithmic curation might not place users in a filter bubble, their own preferences might be causing them to create it for themselves. Scaling effects might also play a role. In "Origins of Homophily in an Evolving Social Network," the authors report on a longitudinal study of a large university community using e-mail interactions, demographic data, and class registration data to create a very complete picture of the social network in the university and its evolution.[18] They find that even a mild "local" preference to associate with like others can cause "induced homophily" in the interaction network—the choice with whom to interact becomes constrained; the compounding of these effects creates stark observed homophily at the population level.

One of the main shifts in the media landscape caused by the advent of social media is an explosion in the number of available sources for information. In this environment, the well-established psychological phenomenon of "selective exposure," introduced by Festinger, can have a particularly strong effect. Selective exposure refers to "the phenomenon whereby people choose to focus on information in their environment that is congruent with and confirms their current attitudes in order to avoid or reduce cognitive dissonance."[19] Berkowitz offers a comprehensive review.[20] Both Lewandowsky and Spohr point out that in such high-choice media environments, the combination of finite attention and selective exposure could drive users to engage almost

completely with content that agrees with their prior views, even if they are exposed to cross-cutting content. These fears are realized in the data.[21] In "Anatomy of News Consumption of Facebook," the authors show that, initially, users on Facebook engage with a fairly large variety of pages, but as time goes on they converge to engaging with only a small group.[22] This tendency is evident of users expending effort to find what they like and then sticking to it; in other words, evidence for selective exposure at work. This effect is confirmed in a subsequent study.[23] Additionally, both of these studies find that, across all users, the most active users focused their attention on the fewest number of pages, a finding that Cinelli and colleagues report as well.[24]

Barberá and colleagues were early adopters of the idea that selective exposure could be a main driver of the emergence of homophily in social media. They pose a successful ideological scaling model that frames the following of a politician or news source on social media as a costly action (capturing finite attention) and encodes a preference for following those whose ideology values are similar to one's own (encoding homophily in the following network). The assumption that the structure of the observed following network was driven by users choosing to follow other users who were close in ideology resulted in fitted ideology values that exactly replicated ideology values for politicians computed from voting records and values for ordinary citizens computed from self-reporting on Twitter profiles, campaign contribution records, and voter registration records independently.[25] Selective exposure is not a tendency induced by social media usage; there is strong evidence of its effects in traditional media consumption outside social media as well.[26] This psychological tendency, in combination with the scale of information available on social media, however, is likely a driver of the striking degree of homophily observed in social networks. All the empirical research into selective exposure on social media illuminates the existence of large, homophilic clusters of users who selectively expose themselves to information that is often verifiably false.[27]

Why would users prefer to expose themselves to information that is falsifiable? Kahan offers an explanation echoed by Spohr[28]; this preference is likely a mechanism to establish and maintain group identity. Far from being irrational, Kahan argues, the propensity to ignore facts in favor of agreeing with one's group is rational in the sense that any gain resulting from an individual shift in ideological position is far outweighed by the perceived cost from the resulting social backlash and the loss of all the social advantages group membership carries. This idea is borne out empirically in social media data as well. Zollo and colleagues exposed users who had a preference for content produced by conspiracy pages to a variety of debunking content. Conspiracy users very rarely engaged with such content. When they did, the action resulted in their becoming more polarized toward the conspiracy camp.[29] Garimella and colleagues find that the community punishes users on Twitter who are "bipartisan" in the sense that they both consume and share content across the ideological spectrum in question. These users have lower network centrality and lower engagement. Conversely, the narrower in ideological scope a user's produced content was (given sufficient activity), the more engagement they received.[30]

Effects of Homophily

The 2016 election launched concerns that social media might act as an amplifier of misinformation into the mainstream. There is substantial evidence that these concerns are justified. Allcott and colleagues analyze a collection of 156 false news stories circulated on Facebook concerning the 2016 US presidential election. These 156 stories were shared a combined 37.9 million times. They estimate that every US adult on Facebook, on average, saw and remembered at least one false story. They hint at the role of homophily in the prolific spread of false news, by showing through survey work that both Democrats and Republicans are 14 percent more likely to believe news that is

ideologically congruent, and that a strong correlate with ideologically aligned inference is self-reporting of a large share of Facebook friends having the same political ideology.[31] Research leveraging large-scale social media data brings the role of homophily to center stage.

As noted, users on social media organize themselves into homophilic clusters with respect to ideology on many issues. This formation appears to be driven, at least in part, by selective exposure, and users in these clusters strongly favor ideologically aligned content. In light of these facts, it is reasonable to suppose that homophily might contribute to the spread of misinformation on social media. Modeling of information diffusion processes on real social media network structures appear to support this hypothesis. Cota and colleagues identify the presence of strong homophily in the network structure surrounding the impeachment of then Brazilian president Dilma Rousseff.[32] In order to examine how homophily might affect information spread in the network, they simulate slightly modified versions of the classical epidemiological susceptible-infected-susceptible (SIS) and susceptible-infected-recovered (SIR) models using the actual measured social network as the contact network for the models.[33] They find that information diffusion on the homophilic network is biased toward individuals who share the same political opinion; given a user and a piece of content received by that user, it is very likely that said content originated with another user of the same political leaning. This same methodology is extended to a much wider range of issues and platforms with confirmed homophily by Cinelli and colleagues with similar results; in all issues where a homophilic network was observed, users with a given ideology are much more likely to be reached by information that is spread by users with a similar ideology.[34]

One might be tempted to argue that the ease with which information propagates within homophilic network components (and the symmetric difficulty that it has propagating between them) is merely a structural consequence of the network topology and doesn't necessarily imply that homophily aids the spread of misinformation. Though network structure surely plays a role, this argument ignores the mechanisms that likely drive the formation of homophilic network structures in the first place. As discussed, there is much evidence that people prefer information that is aligned with their ideology even when this information is verifiably false. Thus, it is only reasonable to expect that false information could widely propagate in a homophilic network component if it aligns with the ideology of the component.

For example, Del Vicario and colleagues find that homophily is the main mechanism behind the effective spread of content. They analyze content cascades on Facebook relating to science news and conspiracy news (derived from the same large dataset studied in "Echo Chambers on Facebook").[35] A content cascade is the successive sharing of a piece of content allowing it to spread through a social network. One can think of a cascade as a tree structure branching through the network, and rooted at the user who originally posted the content. Del Vicario and colleagues define the polarization of a user with respect to science and conspiracy as a value between -1 and 1 via the link-to-labels strategy using a list of labeled science and conspiracy pages; a user with polarization -1 likes only science-related pages, a user with polarization 1 likes only conspiracy-related pages.[36] They then define the *edge homogeneity* of a friendship link between users as the product of the user's polarization values: edge homogeneity is positive when two users have the same ideology and negative when their ideologies differ. This method allows them to study the role of homophily in cascade dynamics by examining the average edge homogeneity of viral cascades. Strikingly, they find that the average edge homogeneity of a cascade is *always* significantly positive. Science-related content is circulated only by users polarized toward science- and conspiracy-related content is circulated only by users polarized toward conspiracy.

These findings suggest that homophily is necessary for large viral cascades to occur; in a

sufficiently sized heterophilous network, it would be rare to find the multitude of positive-edge homogeneity paths necessary to facilitate a viral cascade. The more homophilic a network is, the larger the size of the cascades it can facilitate. The research also suggests that in addition to being seeded in a homophilic network component, content must align with the ideology of the component in order to initiate a cascade of shares. Though Del Vicario and colleagues do not attempt to verify any of the content in the cascades they study, they do find that conspiracy cascades are on average much larger than science cascades. If one makes the (not-so-strong) assumption that the conspiracy set of content contained more false information, one could infer that false information spreads more effectively.

Vosoughi, Roy, and Aral make this inference concrete. They studied the structure of viral cascades on Twitter having content that was verified by independent fact checkers as true or false. They found that false information spreads much more effectively than truth and that this spread is driven by individuals:

> Whereas the truth rarely diffused to more than 1000 people, the top 1% of false-news cascades routinely diffused to between 1000 and 100,000 people. Falsehood reached more people at every depth of a cascade than the truth, meaning that many more people retweeted falsehood than they did the truth. The spread of falsehood was aided by its virality, meaning that falsehood did not simply spread through broadcast dynamics but rather through peer-to-peer diffusion characterized by a viral branching process. It took the truth about six times as long as falsehood to reach 1500 people and 20 times as long as falsehood to reach a cascade depth of 10. As the truth never diffused beyond a depth of 10, we saw that falsehood reached a depth of 19 nearly 10 times faster than the truth reached a depth of 10. Falsehood also diffused significantly more broadly and was retweeted by more unique users than the truth at every cascade depth.[37]

Combined with the insights of Del Vicario and colleagues, this finding suggests that the prolificness of false news on social media can be explained by two factors that are likely working in concert: false news is seeded in more homophilic network components than the truth or it is more aligned with the ideology of the network component it is seeded in than the truth or both.

Though the wide spread of misinformation on social media facilitated by homophily can hardly be disputed, some have questioned its real world effects (especially in elections).[38] Research into the causal effects and mechanisms of how information received on social media affects real-world decision-making is in its infancy. But as an example case study, social media data on the topic of vaccines presents some striking correlations. A 2020 report by the Center for Countering Digital Hate found that social media accounts held by anti-vaccine advocates had increased their following by at least 7.8 million people since 2019. At that time, 31 million people followed anti-vaccine groups on Facebook, with 17 million people subscribing to similar accounts on YouTube.[39] These communities are highly homophilic, and homophily appears to be a main driver in the spread of the misinformation that defines these groups.[40] Cinelli and colleagues, for example, show that the entire platform Gab is a homophilic cluster with respect to vaccine stance and subsequently demonstrate that vaccine misinformation spread most effectively on Gab.[41]

Perhaps most worrying, Johnson and colleagues demonstrate that anti-vaccine clusters are "winning" on Facebook despite their smaller overall size when compared with pro-vaccine clusters. Anti-vaccine clusters occupied more central network positions and were more entangled with each other (homophily) and with clusters that were identified as being undecided with respect to vaccines. This last point shows that anti-vaccine clusters have a greater opportunity to spread their views to undecided clusters.[42]

A modeling effort undertaken in the same study suggests that if the current dynamic remains unchanged, then anti-vaccine discourse could dominate the conversation within the next decade. Spohr notes that for users in homophilic communities where misinformation is spreading (such as anti-vaccine communities), *availability bias* would likely impact their decision making.[43] Availability bias refers to the tendency of humans to over-rely on information that is easily recalled when making a decision. Since more recent information is more easily recalled, this over-reliance results in a heavier weighting of recent information in the decision-making process.[44] Spohr points out that if misinformation is spreading in a homophilic community, then users in that community are likely to make real-world decisions based on that misinformation as they receive all their information (and therefore all recent information) from within the community. At the time of writing, the CDC estimates that 32.8 percent of eligible Americans were not fully vaccinated against COVID-19.

Sunstein gives a stark warning about one of the possible malign effects of homophily in "The Law of Group Polarization."[45] He notes that with marked regularity, empirical studies find that group deliberation results in the group polarizing in the direction of the initial group tendency: a group of vaccine skeptics is likely to be even more skeptical after conferring with one another, those in favor of stringent gun control will likely favor even stronger restrictions after discussion, those who entertain the possibility that the earth is flat will more strongly subscribe to the belief after interacting with each other, and so on. Sunstein calls this phenomenon "the law of group polarization." Because of the pronounced homophily in social media networks, the law of group polarization does not paint an optimistic picture with respect to social media's ability to foster healthy deliberation in society. Instead, it predicts that continued social media interaction in homophilic networks with respect to a given issue will increase polarization in the population on that issue.

The picture becomes gloomier when considering the law of group polarization in the context of social media's homophily-driven vulnerability to misinformation. Sunstein notes that the process through which group polarization from deliberation occurs likely shares many qualities of informational cascades. Misinformation could exacerbate the size and speed of such "polarization cascades." Consider, for example, a homophilic network component of anti-vaccine users, where one of the more extreme users begins espousing the (untrue) claim that vaccines are really intended to insert a government-tracked microchip and should thus be avoided. Several other, more extreme, users adopt the claim on the basis of its congruence with their ideology alone. Sunstein notes that one of the mechanisms of the law of group polarization is that often the most extreme members of a group have the most persuasive power at their disposal. This notion is supported by social media data; as we noted earlier, activity, engagement, and network centrality all correlate with ideological extremity. Next, through persuasive arguments rooted in ideology, the extreme initial adopters of the government-control theory successfully persuade a plurality of their more moderate (but still anti-vaccine) compatriots. At this point, cascade effects become prominent.

Sunstein points out that the beliefs that others espouse carry an informational externality about what it makes sense to believe. Following the adoption of the government-control view by convinced moderates, those individuals most initially hesitant to the view (but still anti-vaccine) will likely adopt it as they are incentivized both by its congruence with their existing ideology (vaccines are bad) and by the fact that many of their peers have adopted it. Finally, once the view is adopted by the whole group it can be used as the basis of further persuasive arguments and be accepted as a known fact.

At the end of this process not only has the anti-vaccine community become more polarized toward anti-vaccine views (because they also now believe them to be a government conspiracy),

they have incorporated a piece of misinformation into the epistemology of their community, thereby making it more susceptible to further polarization. Sunstein notes that the other likely mechanism underlying the tendency of deliberating homophilic groups to polarize is the existence of a skewed argument pool. Since the deliberating group shares an ideology, most of the persuasive arguments available to the group skew in the direction of the shared ideology—repetition of these arguments among the group further polarizes the group. Adding ideology-congruent misinformation to this dynamic has the effect of expanding the pro-ideology argument pool, causing further polarization. This is perhaps one of the most dangerous possible effects of homophily in our current information landscape. As the polarization process iterates, homophilic communities are susceptible to incorporating more and more misinformation to their joint conception of what is known and knowable.

Lewandowsky also worries about the creation of alternative epistemologies in communities where misinformation readily spreads. He writes:

> The framing of the current post-truth malaise as "misinformation" that can be corrected or debunked fails to capture the full scope of the problem. This framing at least tacitly implies that misinformation is a blemish on the information landscape—our mirror of reality—that can be cleared up with a suitable corrective disinfectant. This framing fails to capture the current state of public discourse: the post-truth problem is not a blemish on the mirror. The problem is that the mirror is a window into an alternative reality.[46]

When a homophilic community has constructed an understanding of the world for itself that is based (at least in part) on ideologically aligned misinformation that spreads throughout the group, the argument can be made that members within the community have taken steps toward experiencing *identity fusion* with the group. Identity fusion occurs when an individual's personal self (characteristics that make someone a unique person) and their social self (characteristics that align the person with certain groups) become joined; individuals who have experienced identity fusion feel a profound "oneness" with the group.[47] Identity-fusion measures have been shown to predict an individual's willingness to fight and die on the group's behalf.[48] At this point homophily-driven echo chambers are not just narrowing information landscapes and degrading discourse, they are breeding extremism.

Modeling Homophily through Bounded Confidence

As previously discussed, one of the major effects of the introduction of social media to the information environment is the removal of intermediaries in informational sources, as well as the creation of the possibility of direct interaction with a larger group of individuals. Thus, the process of opinion formation on social media can be seen as an example of *self-organized dynamics*: large-scale behaviors emerge without a central authority, much like the behavior of a flock of birds or a shoal of fish. To study how small-scale individual interactions can result in large-scale structure like the observed homophily in social media, agent-based models of opinion formation are often employed. Here, an individual's opinion is modeled as a continuous value on a one-dimensional spectrum that changes in accordance with that individual's connections with other individuals in a social network modeled as a directed graph. A given individual feels a "push" or "pull" exerted on their opinion value by those who interact with it (point at it in the directed graph) according to prescribed interaction rules.

Most models in the literature have an interaction rule that encodes an assumption of "local consensus." In the absence of other interactions, when one individual interacts with another, the interaction exerts a pull on the second individual's opinion until that second individual's opinion

is the same as the first one's. When two individuals interact, they exert a mutual pull on each other's opinion and will eventually reach a consensus in the middle of their original opinions. Often, the structure of the social network changes depending on the distribution of opinions in the population. For example, two individuals might begin interacting/cease interacting (resulting in an edge being added/deleted in the network) if their opinion values become close enough together/too far apart. The entire collection of opinions is then allowed to evolve until it reaches (or does not reach) an equilibrium state.

A hallmark of the study of these models is examining how the interplay between the topology of the underlying network and the interaction rules affect the distribution of opinions among the agents. Much of the mathematical literature exploring such models is concerned with studying conditions that cause evolution to a *consensus* (all agents have the same opinion value) in equilibrium. As with much mathematical analysis, simpler cases prove very instructive. By examining the case where the structure of the social network remains unchanged throughout the evolution of the model, it has been shown that a necessary condition for the emergence of a consensus (in the case of an attractive interaction rule) is the persistence of a suitable degree of connectivity in the network.[49] This allows for *heterophilic* interactions: agents with disparate opinions interact and because of the attractive nature of the interaction rule, eventually agree.

One might assume that interaction rules carrying a local consensus assumption cause the emergence of a global consensus to be a ubiquitous feature; but it does not. The manner in which agents are connected in the underlying network has a large effect on the distribution of opinions observed among the agents. One of the more well-studied models, known as the "bounded-confidence model," was introduced by Hegselmann and Krause.[50] Here, the connections between agents are dynamic: a connection forms between agents when their opinions are within an interaction range. If two agents have opinions within the interaction range, they attract each other; otherwise they feel no influence from each other. This dynamic causes the formation of "clusters" of opinions in the longtime limit to be a generic behavior; consensus is rare. For this reason, much of the study of this class of models has focused on analytically characterizing the clustering behavior.[51]

The interaction range in bounded-confidence dynamics causes the underlying social network to be homophilic; agents interact only with agents who are sufficiently similar in ideology. This tendency causes the social network to quickly become disconnected into a collection of echo chambers that share ideology, preventing a consensus from occurring even though agents who do interact attract each other. The bounded-confidence interaction rule encodes the previously discussed tendency of selective exposure; agents interact and agree with those who are close to them in ideology and ignore the rest. The interaction networks generated by bounded confidence type models have been seen in many cases to replicate the homophilic interaction patterns seen in real social media data.[52]

We employ the bounded-confidence interaction rule to examine how populations that are prone to selective exposure are susceptible to misinformation. To do this, we introduce the notion of a *polarizing agent*. In a normative social network, agents both feel and exert influence. A polarizing agent exerts influence only on those to whom it is connected in the social network and does not adjust its opinion value as a result of its interactions. This is also the role of misinformation in real social networks. To illustrate the effects of polarizing agents on homophilic social networks, we perform a series of simulations of the bounded confidence dynamics in the presence of polarizing agents and without them (see Figure 1). In the top plot of Figure 1 we simulate the bounded confidence dynamics among a population of two hundred agents without the presence of polarizing agents. The selective exposure mechanism encoded in the bounded-confidence

interaction rule causes the underlying social network to become homophilic, and two clusters of agents emerge (corresponding to two homophilic components of the social network); notice, however, that the variance in opinion of the entire collection of agents decreases over the evolution of the model. In the bottom plot of Figure 1 we repeat the simulation from the same initial condition and add polarizing agents (shown in red) at opinion values 1.6 and -1.6. In this case, two clusters of agents also emerge; however every agent in each cluster eventually takes on the opinion of the polarizing agents closest to its cluster. The clusters have much more separation in ideology in the presence of polarizing agents. This increased susceptibility to polarization is reflected in the variance in opinion of the population, which increases over the entire evolution in contrast to the case without polarizing agents.

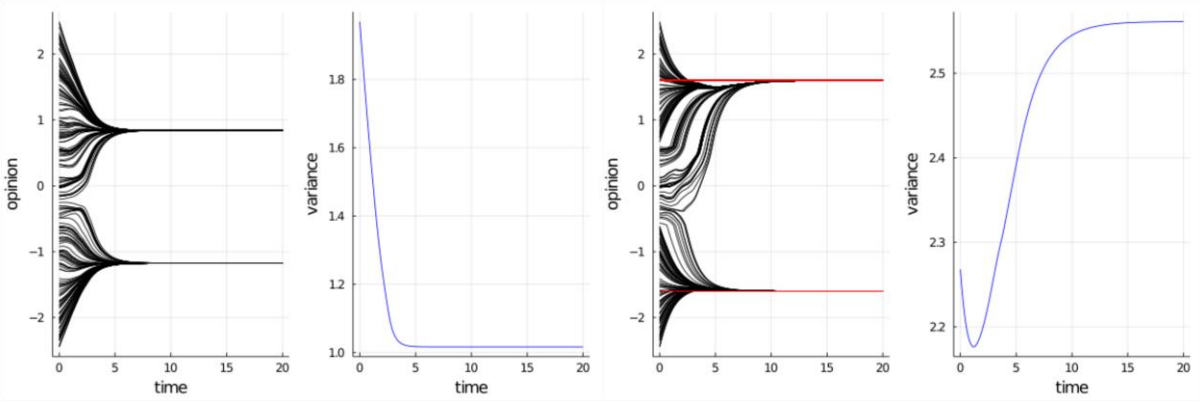

Figure 1. In the top plot, the bounded-confidence dynamics are simulated with 200 agents without the presence of polarizing agents. In the bottom plot the dynamics are simulated from the same initial state in the presence of polarizing agents (shown in red) at opinion values 1.6 and -1.6. In the top simulation, homophily in the interaction network causes the formation of two clusters of agents however the variance of the overall collection of agents decreases. In the presence of polarizing agents (shown in red), the variance increases over the course of the simulation, indicating the susceptibility of the homophilic clusters to further polarization.

It is possible to induce consensus in a population of agents prone to homophily by imbuing some of the agents with a moderating tendency. In keeping with the bounded-confidence interaction rule, cluster formation is equivalent to a fragmentation of the interaction network into homophilic components. If the connectivity of the interaction network is maintained, the attractive nature of the interaction rule should result in a consensus. It has been mathematically proven that if there are several moderating agents that do not adjust their opinion value if the interaction network is close to severing, then a consensus is reached in a population otherwise evolving according to the bounded-confidence rule.[53]

In Figure 2 we examine the effects of polarizing agents on a population that includes moderating agents evolving according to the bounded-confidence interaction. In the top two plots, the bounded-confidence dynamics are simulated in a population that includes only moderating agents. The right top plot shows the entire evolution and the left top plot shows the first twenty time units to give a better sense of the role of the moderating agents. Since the moderating agents do not adjust their opinions when the interaction network is close to fragmenting and begin evolving again only when this "danger" has passed, the connectivity of the interaction network is maintained throughout the evolution. The original two homophilic clusters do form initially (as well as a new moderate one); but the moderating agents cause these clusters to eventually merge

and a consensus is reached. In the bottom plot we perform the same simulation in the presence of polarizing agents. Though a new moderate cluster does emerge, the majority of the population still become polarized, as evidenced by the increasing variance. Since polarizing agents never need to adjust their opinions to achieve their "goal," and moderating agents do, an impasse is reached. Connectivity of the interaction network is maintained; but most of the agents become polarized and the moderate regions of the ideology spectrum are populated mostly by moderating agents and the small new moderate cluster.

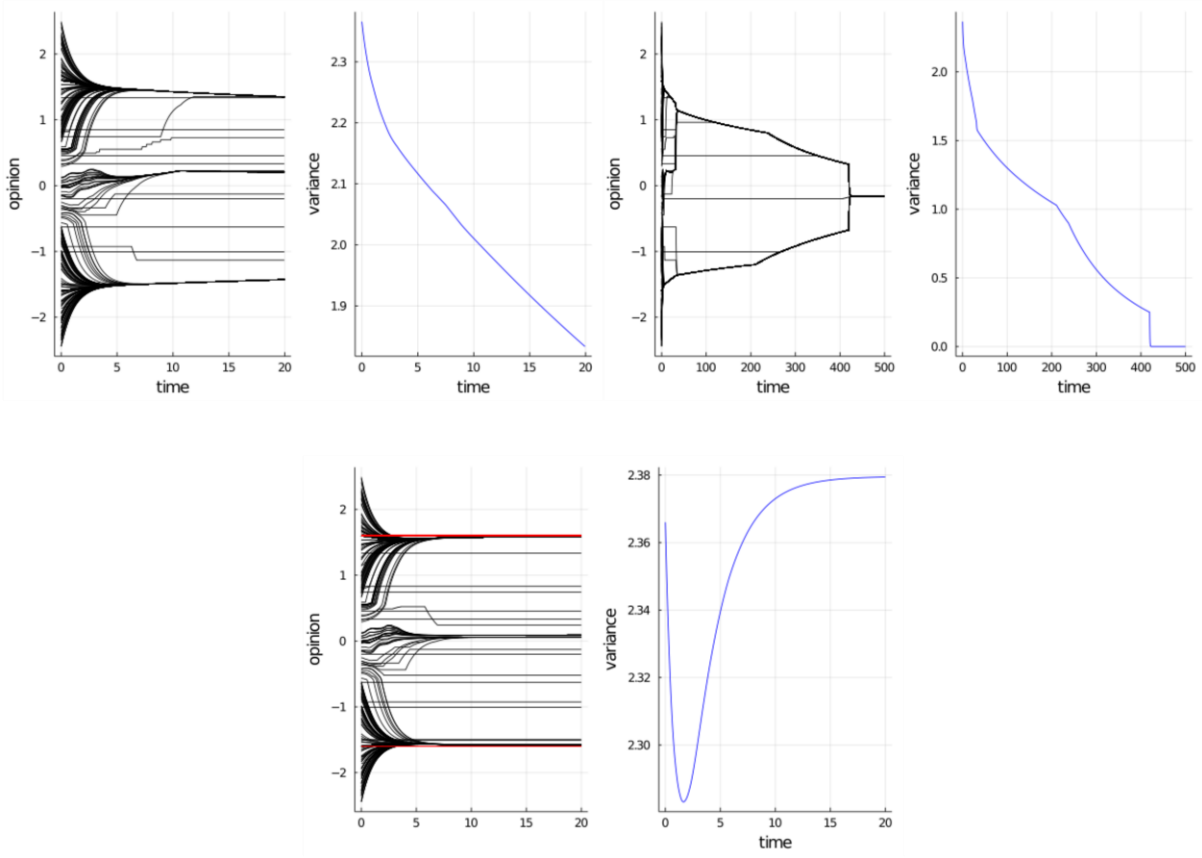

Figure 2. In the top two plots, the bounded-confidence dynamics are simulated in a population that includes moderating agents but not polarizing agents. The top-right plot shows the entire evolution and the top-left plot shows the first 20 time units to give a better intuition as to the role of the moderators. The moderating agents maintain the connectivity of the interaction network and a consensus is reached. In the bottom plot the same simulation is repeated but with the addition of polarizing agents. The moderating agents do result in the formation of a moderate cluster but the polarizing agents succeed in polarizing the majority of the population as indicated by the increasing variance.

These theoretical results suggest that moderating from the middle is not sufficient to prevent polarization in a population prone to homophily if polarizing agents (misinformation) are present. In this setting, it seems that the only hope for consensus is to "cut off" the polarizing agents from the population and allow the moderating agents to depolarize the population without competition from the polarizers. As discussed earlier, it has been shown in real social media data that moderate or bipartisan users are less successful than their extreme counterparts in several ways and that debunking or moderating messages directed at homophilic communities are usually ignored. Additionally, there is a growing body of evidence that "deplatforming" polarizers is an effective

strategy for mitigating polarization in the rest of the population.[54]

Artificial Intelligence as a Polarizing Agent

Social media's tendency to organize into homophilic clusters and the resulting vulnerability of the clusters to polarization create a landscape ripe for exploitation by malign actors. The narrowing of the information landscape that occurs in a homophilic cluster allows for the values important to such a cluster to be easily characterized. These values can then be wedges that when properly deployed could be used to shift the cluster's opinion on other issues.

To illustrate our point, we return to the example of a conspiracy rumor about government microchipping being seeded in an anti-vaccine cluster. We noted that the spreading of the conspiracy rumor in the anti-vaccine cluster had two consequences for that cluster: it became further polarized toward vaccines and incorporated a piece of misinformation to its epistemology, thus becoming more susceptible to polarization. But there is a third effect: the cluster also became more negatively polarized toward the government. In this example, mistrust of vaccines is the core value leveraged in order to shift the anti-vaccine community's opinion of the government. We claim that recent developments in *natural language processing*, specifically in the techniques of *language modeling*, *topic modeling*, and *sentiment analysis*, create the possibility of an automated capability for influencing homophilic clusters. With some initial data structuring from subject-matter experts, this capability could automatically measure the values important to a cluster and generate original messaging salient to those values that is designed to shift opinion on a different set of *target values*. The social media environment allows for the distribution of such messaging at scale, directly to users in the target cluster. We first present a general, top-down view of the goals of language modeling, topic modeling, and sentiment analysis and then describe how they could be used in concert to achieve the described influence capability.

Language Modeling

Broadly, *natural language processing* aims to leverage statistical modeling of textual data for application to a large variety of automated tasks. Common examples include text classification (Is this e-mail spam or not?), text generation ("Siri, write me a poem!"), topic modeling (What is this large collection of documents concerned with?), and sentiment analysis (Is this sentence positive or negative?), among many others. Before solving these "downstream tasks," one needs to find a way to represent text in a way that preserves a signal of its meaning and can be processed by a computer. The task of finding such a representation is known as language modeling. More concretely, consider a collection of text (often referred to as *the corpus*), T, made up of a vocabulary of words $V = \{ w_1, ..., w_N \}$. For each word, w_i, the aim is to learn a vector, $\tilde{w}_i$, such that words that have a similar meaning have vector representations that are geometrically close. For example, since the word "queen" is more similar to "king" than "orange," a good representation should place the vectors for king and queen closer together than the vectors for queen and orange (see Figure 3).

This begs the question: How do we define statistically what it means for two words to have a similar meaning? To answer this question *the distributive hypothesis*, a concept from linguistics, is leveraged. The distributive hypothesis states that words that have similar meanings appear in similar textual contexts. Under the distributive hypothesis, the problem of learning a good representation becomes as follows: for each word in the vocabulary, learn a vector in such a way that words that appear in similar contexts have vector representations that are geometrically close.

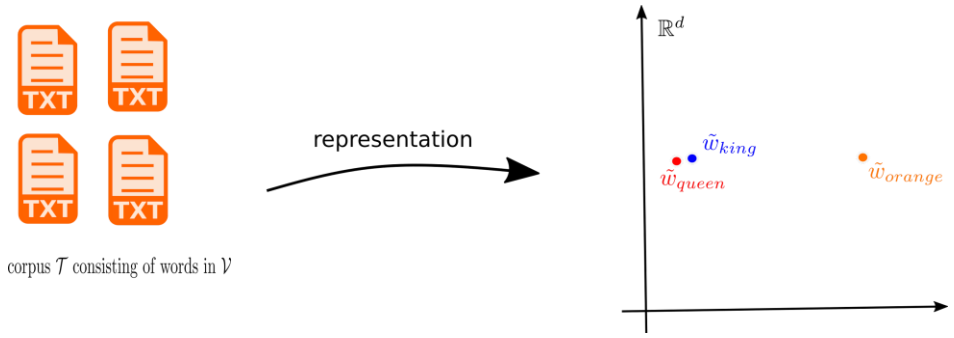

Figure 3. A good language model should place vector representations of similar words geometrically close. In practice, the context of a word is defined as a window of preceding words in the corpus or a window of words on either side (a given word can have multiple contexts).

$C(w_i) = (w_{i-m}, ..., w_{i-1})$ context of the ith word is *m* preceding words

$C(w_i) = (w_{i-m}, ..., w_{i-1}, w_{i+1}, ..., w_{i+m})$ context of the ith word is *m* words on either side

A *training set* of pairs of words and their respective contexts is formed from the corpus, and, with the use of a statistical learning model, the word vectors are fit to the objective of "given a context, predict its word." This is exactly a "fill in the blank" task. Given how the context of a word was defined, this objective is really "given several preceding words, predict the next word" or "given a sequence of words with a missing word, predict the missing word."

Roses are red, violets are _____ → [model predicts] → blue

or

Roses are _____ , violets are blue → [model predicts] → red

The fitted word vectors are retained for use as representing text in the previously mentioned downstream tasks. But in the case that context was defined as preceding words and the model was trained on the "predict the next word" task, the trained model can be prompted successively to *generate* new pieces of text:

Roses → [model predicts] → are

Roses are → [model predicts] → red

Roses are red → [model predicts] → violets

Roses are red violets → [model predicts] → are

Roses are red violets are → [model predicts] → blue.

This generation can even be controlled to an extent if the text data has some structuring done to it before training. For example, if training text examples are prepended with the topics they pertain to, then the model could be controlled to generate text salient to a given topic by prompting it with the topic itself. The language modeling approach is very general; much of the research in natural language processing focuses on the models and techniques used to fit the representation vectors. Currently, the models that perform best on all benchmarks are deep-learning models known as "transformers," which have been trained on large text datasets at the cost of millions of dollars.

Topic Modeling

Topic modeling aims to measure important topics that appear in a large text corpus. Intuitively it aims to give a sense of what a collection of documents is about. Functionally, it is a two-step process. First, important words and phrases in the corpus are identified; for example, common "stop words," such as articles, prepositions, pronouns, and conjunctions, are filtered out. Then, the important words and phrases are clustered into groups that can be (if the model did a good job) interpreted as abstract topics; these groups of phrases are then usually named by a human annotator. For example, if one were modeling a collection of news articles having to do with a treaty process between two countries, some measured topics might be:

agreements = [deal, treaty, peace deal, ceasefire agreement, trade representative, climate agreement, trade commission]

defense = [security, soldiers, base, fighters, militants, weapon, militia, conquest, patriot]

economy = [companies, exchange, capacity, wages, taxpayers, inflation, inequality, banker, oil prices, debts, stock market].

One main use for a completed topic model is to detect the topics that are present in an arbitrary piece of text from the same domain that the topic model was fit to. For example, given an arbitrary news article, we could detect treaty process specific topics using the example topic model above. There are myriad techniques for computing topic models, both leveraging traditional statistical approaches and deep-learning techniques. One major challenge across all techniques is evaluation of the quality of a computed topic model: the most agreed on technique is the use of human evaluators who are subject-matter experts in the domain with which the text corpus is concerned.

Sentiment Analysis

One of the most common downstream uses for language model representations is sentiment analysis. Sentiment analysis aims to classify whether a given piece of text has positive, negative, or neutral sentiment. In its simplest form it assigns one sentiment value to the entire piece of text. For example:

I hate apples → negative

I love bananas → positive

I hate apples and love bananas → neutral.

The final example above illustrates one of the drawbacks of considering the "global" sentiment of a piece of text and is a typical output of such algorithms. A piece of text might contain multiple sentiments that get "averaged" when considering its overall sentiment. Topic modeling provides an avenue to solve this problem. Given a set of topics, sentiment analysis algorithms can be trained to detect the sentiment in a piece of text with respect to each topic. This is known as *aspect-based* or *topic-based* sentiment analysis. For example, if we had a large corpus of sentences concerning food and computed a topic model on it, we might find that two of the topics are "fruits" and "vegetables." We could then use that topic model to detect the topics present in each sentence and label the sentence with its respective topics. If we then trained a topic-based sentiment-analysis algorithm on the topic-labeled sentences, we could then use it to detect that the sentiment toward

fruits in the sentence "I hate apples and love bananas" was positive and that the sentiment toward vegetables in the same sentence was negative.

Artificial Intelligence as a Polarizing Agent

All three of the capabilities outlined in the previous section have seen explosive progress in the past decade. Topic modeling and sentiment analysis are basic tools in the toolkits of any company trying to understand its customers' online behavior. For example, they are the main ingredients in any system meant to make recommendations based on previous behavior. Downstream tasks for language models have such a wide application in industry and academia that entire businesses are now devoted to the expensive pretraining of large language models in order to provide them as services (e.g., HuggingFace and OpenAI). Society has made a sizable investment in these techniques and we should expect only to see their performance increase and their use become more widespread. As with any emerging technology, malign actors have the opportunity to put it to use. Society should expect this threat and craft policy to confront it.

In the previous section we saw from a modeling perspective how homophilic communities are susceptible to influence from polarizing agents. The natural language-processing capabilities outlined can be combined to create an automated system to deploy polarizing agents at scale on social media. In this scenario a malign actor aims to coalesce/fragment a target community's opinion toward/away from a target topic. First, given a target homophilic community, topic modeling is deployed on a large dataset of text scraped from the target community. This topic model is cleaned and validated by human analysts and any topics corresponding to values of the target community are noted. Next, the topic model is used by a subject-matter expert to craft a campaign design that leverages the measured values in order to shift the community's opinion on the target topic. For example, to coalesce the community toward the target topic, automated polarizing agents should inundate the community with messaging that frames community values and the target topic positively, as well as messaging that frames community values and the target topic negatively. To fragment the community away from the target topic, messaging should be pushed that frames the target topic positively and community values negatively as well as messaging that frames the target topic negatively and community values positively. In the hands of analysts knowledgeable about the target community, the topic model could be used to craft more intricate campaign designs.

In parallel, also using the topic model, aspect-based sentiment analysis is deployed on a large collection of text scraped from the target community to tag each text example with its sentiment towards the important values to the target community. This tagged dataset is then used to train a controllable, generative language model. Such a model will be able to speak to the community in its own parlance because it was trained on text data taken from the community, and it will be able to speak both positively and negatively to the community's values as each training example was tagged with its sentiment toward those values (see Figure 4).

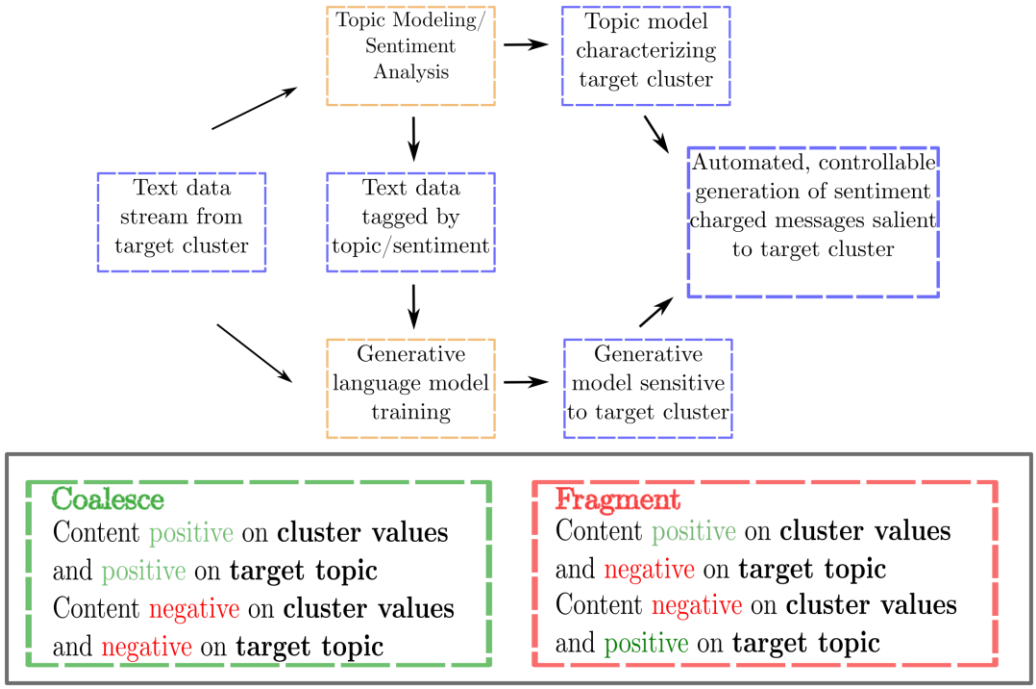

Figure 4. A process to deploy automated polarizing agents. Such agents could be used, for example, to carry out an automated, at-scale campaign aimed to coalesce/fragment a target cluster toward/away from a target topic through leveraging values important to the target cluster.

At this point, all the main components are in place to deploy automated polarizing agents. Sentiment charged messaging salient to the values of the target community can now be generated in an automated and controllable fashion according to subject-matter-expert-crafted rules. This capability and the means of dissemination provided by social media opens the possibility of deploying such messaging at massive scale. Significant technical challenges must be overcome to successfully deploy this strategy, mainly the creation of a large collection of social media accounts to push the messaging through and the corresponding command and control infrastructure to automate the actual posting behavior. A sufficiently funded and motivated actor, however, would surely overcome such hurdles.

At Artis we have observed evidence of such a strategy being deployed in real social media data in ongoing research. In Figure 5 we visualize a Twitter dataset from 293,046 anonymized Twitter accounts that were discussing gun control in the United States. Individual nodes in the network are users, red nodes have been classified as anti-gun-control and blue users as pro-gun-control using an automated content classifier. Edges represent retweets and are colored by the color of the retweeting account. The network clearly has a homophilic structure. Yellow nodes were identified as accounts who displayed statistically abnormal and coordinated posting activity; these accounts amplified the same content in a coordinated manner. While a quantitative analysis of the content pushed by the yellow nodes was not done, qualitative examination revealed that yellow nodes in the anti-gun-control community were pushing content that framed gun-control measures as dangerous to the health of the country; interestingly, yellow nodes in the pro-gun-control community were pushing the same framing except with respect to relaxing gun-control measures. Though we do not have evidence that the campaign messaging itself was automatically generated in the way we describe, the coordination of the yellow nodes' activity does speak to the existence of a sophisticated automated command and control infrastructure. The positioning of yellow nodes

at the centers of homophilic network components is telling as well.

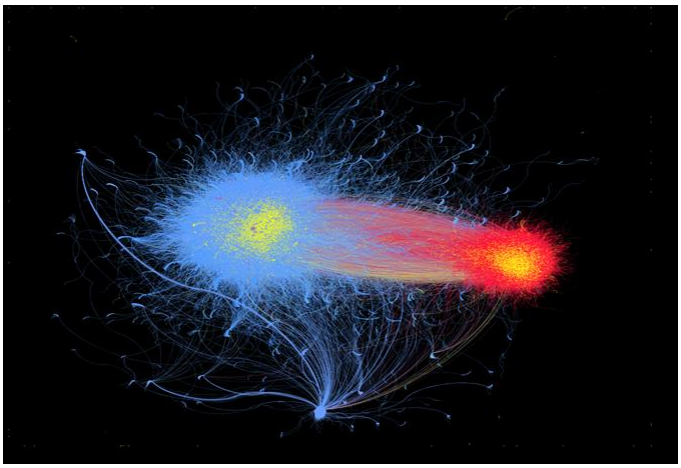

Figure 5. A visualization of a Twitter dataset from 293,046 anonymized Twitter accounts concerning the gun-control debate in the United States. Individual nodes are users, red nodes have been classified as anti-gun-control, and blue users as pro-gun-control using a content classifier. Yellow nodes were identified as accounts who displayed statistically abnormal and coordinated posting activity. Edges represent retweets and are colored by the color of the retweeting account.

Conclusion

Research into how to effectively combat homophily in online discourse is still in its early stages. Most studies have looked into either moderation strategies or deplatforming strategies. Moderation strategies aim to depolarize homophilic clusters through "breaking the echo chamber" in some sense. Visually flagging misinformation, exposing users to cross-cutting content, or explicit debunking content are some of the more popular current approaches. As discussed, however, the psychological mechanisms that drive homophily suggest that such approaches will likely fall on deaf ears, and modeling suggests that they are likely insufficient in the presence of polarizers. In the research we review that looks into moderation strategies, they had little success. Deplatforming strategies aim to identify the most extreme drivers of the discourse in homophilic clusters and cut them off from the network. As we discussed earlier, one of the main drivers of polarization resulting from in-group deliberation is the persuasive power wielded by the group's most extreme members. Deplatforming strategies have seen markedly more early success in the literature, often resulting in the reduction of the extremity of discourse within a cluster and occasionally dissolving a cluster significantly. There are worries that deplatforming does not do enough to depolarize users in a cluster and merely results in a "scattering" of the cluster that could result in its reemergence.

Regardless, the problem remains and crafting any policy that effectively addresses homophily in online discourse will surely demand input from a large intersection of expertise. As a clear first step, continued funding should be directed to studying online interaction so that we may more clearly understand how our discourse on the web impacts our society. While the questions are myriad, we offer several areas where we believe future policy and research should focus attention.

As a first policy step, virtually all of the data needed to measure the social media landscape is currently wholly in the hands of the social platforms and disseminated to researchers and other businesses at the platforms' discretion and for a profit. Though most platforms offer a free tier of data collection for public use, these tiers usually represent only a small sample of available data

and the sampling methods are opaque. One consequence of this siloing of data is the dearth of longitudinal analyses of social media data in the literature. Most studies work with static data and do not examine the dynamics of the network structure. To facilitate research into the dynamics of homophily (and other relevant phenomena), platforms should be obligated to make a prescribed amount and type of data publicly available; the extent of this obligation should be tied to the extent that use of the platform has been adopted by the population. Such a policy must balance many competing interests. Platforms would surely argue that much of their profitability comes, for example, from their analysis of proprietary data. But the extent to which homophily presents on social media clearly demonstrates that society needs a clearer picture of itself in the online space.

On the research side we believe that there are currently two main avenues of inquiry that demand attention: exploration of strategies to reduce homophily that has already presented and further research into the mechanisms that facilitate homophily in order to prevent its emergence in the first place. In the first case, because of its early success, more research into the effects of deplatforming should be conducted. Here, longitudinal data is important because the relevant questions concern what happens to the deplatformed entity and the remaining population in the cluster in the period following deplatforming. We also believe, despite the considerable headwinds, that moderation strategies should continue to be investigated. Most of the moderation strategies in the literature have relied on appeals to facts or logic; such an approach is at best neglecting the shared beliefs that coalesce a homophilic cluster and at worst directly contradicting them. As we've discussed, for messaging to resonate in a homophilic cluster, it must be congruent with the ideals of that cluster in some sense. Any strategy that aims to reduce homophily, whether by moderation or other means, must confront this fact. If a moderation strategy proves successful, especially one that relies on supplying cross-cutting content, a regulatory framework in the spirit of the Federal Communications Commission's now defunct fairness doctrine might be erected around it.

As far as mechanisms go, we claim that there are two important types: individual mechanisms (e.g., selective exposure) and collective mechanisms (e.g., induced structural homophily). Much of the previous research into individual mechanisms has been in an offline context. Thus, it is important to understand how the speed and scale of the online space interacts with previously well-established tendencies such as selective exposure and confirmation bias. Additionally, it is important to understand the effects that incorporation into a homophilic cluster have on the individual. There are multiple case studies of online activity contributing to radicalization: Can an individual become identity-fused with an online community? With regard to collective mechanisms, more work into the dynamics of homophilic clusters, leveraging longitudinal data, is needed. How do homophilic clusters form? Does online interaction merely provide an efficient means to find like others and optimize our preference for congruent information or does it play a causal role in the formation of polarized ideals? Once clusters are formed do they ever fragment? Under what conditions? Most important, given two disparate clusters, do they ever merge? Under what conditions?

According to a 2021 Pew Research Poll, 69 percent of Americans use Facebook, three-quarters of those users say they use it every day, and more than one-half of Americans say they at least sometimes receive their news from social media. This scale of use, the tendency of social platforms toward homophily, and the susceptibility to misinformation that that structure induces (among other ill effects) calls for a commitment by policy makers to engage more fully with this problem. The situation is made more urgent by the possibility of the automated capability we describe. The Internet has democratized access to information in many ways and any policy that effectively addresses the homophily problem on social media must do so in a manner that aims to preserve the ideals of freedom of information and speech. The prevalence of homophily on social media,

however, reveals much about the tension between these ideals.

A century ago, maximally free markets and the human tendency to profit maximize led to the spontaneous development of monopolies and the massive restriction of economic mobility for most, which, in turn, led to unprecedented regulation of the freedom of the markets in order to increase economic mobility and a furious debate about the tension between individual and market freedoms that continues today. Today, maximally disintermediated discourse on social media and the human tendency to selectively expose has resulted in the spontaneous development of homophily in online discourse, which, as noted, has resulted in a restriction in the information landscapes of many users. This restriction, too, calls for regulation of the discourse on social media in order to maximize users' freedom of information and inquiry.

The need for regulated freedoms, and specifically regulated discourse, in order to maximize the health of open society has been acknowledged since the founding of the United States. James Madison, for example, believed that the emergence of factions (i.e., homophily) in society was inevitable. In *Federalist Papers 10*, Madison lays out a series of regulations for how discourse should be conducted in the legislature to minimize the ill effects of homophily. The ill effects of homophily in online discourse call for similar regulation. But the unprecedented scale and speed of online discourse necessitates a deep examination into what form such regulation should take, lest it damage the societal good that online speech contributes. In the words of James Madison:

> Liberty is to faction what air is to fire, an aliment without which it instantly expires. But it could not be less folly to abolish liberty, which is essential to political life, because it nourishes faction, than it would be to wish the annihilation of air, which is essential to animal life, because it imparts to fire its destructive agency.[55]

Notes

[1] Howard Rheingold, *The Virtual Community*, rev. ed., *Homesteading on the Electronic Frontier* (Cambridge: MIT Press, 2000); Robert H Anderson, Tora K. Bikson, Sally Ann Law, and Bridger M. Mitchell, "Universal Access to email: Feasibility and Societal Implications," in *The Digital Divide: Facing a Crisis or Creating a Myth?*, ed. Benjamin S. Compaine, 243–262 (Cambridge: MIT Press, 2001); James S. Fishkin, "Virtual Democratic Possibilities: Prospects for Internet Democracy," paper prepared for conference on Internet, Democracy, and Public Goods, Belo Horizonte, Brazil, November 6–10; Vincent Price and Joseph N. Cappella, "Online Deliberation and Its Influence: The Electronic Dialogue Project in Campaign 2000," *It & Society* 1, no. 1 (2002): 303–329.

[2] Jodi Dean, "Why the Net Is Not a Public Sphere," *Constellations* 10, no. 1 (2003): 95–112; Lincoln Dahlberg, "Cyberspace and the Public Sphere: Exploring the Democratic Potential of the Net," *Convergence* 4, no. 1 (1998): 70–84; Hubertus Buchstein, "Bytes That Bite: The Internet and Deliberative Democracy," *Constellations* 4, no. 2 (1997): 248–263; Eli Pariser, *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think* (London: Penguin, 2011).

[3] John Kelly, Danyel Fisher, and Marc A. Smith, "Debate, Division, and Diversity: Political Discourse Networks in USENET Newsgroups," in Online Deliberation Conference, Stanford University, 2005, 4–3; Cass R Sunstein, "Democracy and Filtering," *Communications of the ACM* 47, no. 12 (2004): 57–59; Cass R. Sunstein, *Republic.Com* (Princeton: Princeton University Press, 2001).

[4] Pablo Barberá, "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data," *Political Analysis* 23, no. 1 (2015): 76–91.

[5] Pablo Barberá, John T. Jost, Jonathan Nagler, Joshua A. Ticker, and Richard Bonneau, "Tweeting from Left to Right: Is Online Political Communication More Than an Echo Chamber?," *Psychological Science* 26, no. 10 (October 2015): 1531–1542.

[6] Kiran Garimella, Aristides Gionis, Gianmarco De Francisci Morales, and Michael Mathioudakis, "The Effect of Collective Attention on Controversial Debates on Social Media," in *Proceedings of the 2017 ACM on Web Science Conference* (Troy, NY: ACM, June 2017), 43–52; Kiran Garimella, Aristides Gionis, Gianmarco De Francisci

Morales, and Michael Mathioudakis, "Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship," in *Proceedings of the 2018 World Wide Web Conference* (Geneva: International World Wide Web Conferences Steering Committee, April 2018), 913–922; Barberá, "Birds of the Same Feather."
[7] Barberá et al., "Tweeting from Left to Right."
[8] Matteo Cinelli et al., "The Echo Chamber Effect on Social Media, *Proceedings of the National Academy of Sciences* 118, no. 9 (March 2021).
[9] Bjarke Mønsted and Sune Lehmann, "Characterizing Polarization in Online Vaccine Discourse: A Large-Scale Study," *PLOS ONE* 17, no. 2 (February 2022); Ana Lucía Schmidt et al., "Polarization of the Vaccination Debate on Facebook," *Vaccine* 36, no. 25 (June 2018): 3606–3612.
[10] Eytan Bakshy, Solomon Messing, and Lada A. Adamic, "Exposure to Ideologically Diverse News and Opinion on Facebook," *Science* 348, no. 6239 (June 2015): 1130–1132; Robert Bond and Solomon Messing, "Quantifying Social Media's Political Space: Estimating Ideology from Publicly Revealed Preferences on Facebook," *American Political Science Review* 109, no. 1 (2015); Jisun An et al., "Sharing Political News: The Balancing Act of Intimacy and Socialization in Selective Exposure," *EPJ Data Science* 3, no. 1 (December 2014).
[11] Walter Quattrociocchi, Antonio Scala, and Cass R. Sunstein, *Echo Chambers on Facebook*, SSRN Scholarly Paper ID 2795110 (Rochester, NY: Social Science Research Network, June 2016); Fabiana Zollo et al., "Debunking in a World of Tribes," *PLOS ONE* 12, no. 7 (July 2017); Alessandro Bessi et al., "Homophily and Polarization in the Age of Misinformation," *European Physical Journal Special Topics* 225, no. 10 (October 2016): 2047–2059.
[12] Ana Lucía Schmidt et al., "Anatomy of News Consumption on Facebook," *Proceedings of the National Academy of Sciences* 114, no. 12 (March 2017): 3035–3039.
[13] Max Grömping, "'Echo Chambers': Partisan Facebook Groups during the 2014 Thai Election," *Asia Pacific Media Educator* 24, no. 1 (June 2014): 39–59; Wesley Cota et al., "Quantifying Echo Chamber Effects in Information Spreading Over Political Communication Networks," *EPJ Data Science* 8, no. 1 (December 2019): 1–13; Barberá, "Birds of the Same Feather Tweet Together."
[14] Dominic Spohr, "Fake News and Ideological Polarization: Filter Bubbles and Selective Exposure on Social Media," *Business Information Review* 34, no. 3 (September 2017): 150–160; Pariser, *Filter Bubble*.
[15] Elizabeth Dubois and Grant Blank, "The Echo Chamber Is Overstated: The Moderating Effect of Political Interest and Diverse Media," *Information, Communication & Society* 21, no. 5 (May 2018): 729–745; Matthew Gentzkow and Jesse M. Shapiro, "Ideological Segregation Online and Offline," *Quarterly Journal of Economics* 126, no. 4 (November 2011): 1799–1839; Seth Flaxman, Sharad Goel, and Justin M. Rao, "Filter Bubbles, Echo Chambers, and Online News Consumption," *Public Opinion Quarterly* 80, no. S1 (January 2016): 298–320; Soroush Vosoughi, Deb Roy, and Sinan Aral, "The Spread of True and False News Online," *Science* 359, no. 6380 (March 2018): 1146–1151.
[16] R. Kelly Garrett, "The "Echo Chamber" Distraction: Disinformation Campaigns Are the Problem, Not Audience Fragmentation," *Journal of Applied Research in Memory and Cognition* 6, no. 4 (December 2017): 370–376.
[17] Bakshy, Messing, and Adamic, "Exposure to Ideologically Diverse News."
[18] Gueorgi Kossinets and Duncan J. Watts, "Origins of Homophily in an Evolving Social Network," *American Journal of Sociology* 115, no. 2 (September 2009): 405–450.
[19] Leon Festinger, *A Theory of Cognitive Dissonance*, vol. 2 (Redwood City, CA: Stanford University Press, 1957).
[20] Dieter Frey, "Recent Research on Selective Exposure to Information," in *Advances in Experimental Social Psychology*, ed. Leonard Berkowitz (Cambridge, MA: Academic Press, 1986), 19:41–80, https://www.sciencedirect.com/science/article/pii/S0065260108602129.
[21] Stephan Lewandowsky, Ullrich K. H. Ecker, and John Cook, "Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era," *Journal of Applied Research in Memory and Cognition* 6, no. 4 (December 2017): 353–369; Spohr, "Fake news and Ideological Polarization."
[22] Schmidt et al., "Anatomy of News Consumption on Facebook."
[23] Schmidt et al., "Polarization of the Vaccination Debate on Facebook."
[24] Matteo Cinelli et al., "Selective Exposure Shapes the Facebook News Diet," *PLOS ONE* 15, no. 3 (March 2020): e0229129.
[25] Barberá, "Birds of the Same Feather."
[26] Natalie Jomini Stroud, "Media Use and Political Predispositions: Revisiting the Concept of Selective Exposure," *Political Behavior* 30, no. 3 (September 2008): 341–366.
[27] Mønsted and Lehmann, "Characterizing Polarization in Online Vaccine Discourse"; Matteo Cinelli et al., "The COVID-19 Social Media Infodemic," *Scientific Reports* 10, no. 1 (October 2020): 16598; Schmidt et al., "Polarization of the Vaccination Debate on Facebook." Dan M. Kahan, *Ideology, Motivated Reasoning, and Cognitive Reflection: An Experimental Study*, SSRN Scholarly Paper 2182588 (Rochester, NY: Social Science

Research Network, November 2012); Dan M. Kahan, "Climate-Science Communication and the Measurement Problem," *Political Psychology* 36, no. S1 (2015): 1–43.
[28] Kahan, *Ideology, Motivated Reasoning, and Cognitive Reflection;* Spohr, "Fake News and Ideological Polarization."
[29] Zollo et al., "Debunking in a World of Tribes."
[30] Garimella et al., "Political Discourse on Social Media."
[31] Hunt Allcott and Matthew Gentzkow, "Social Media and Fake News in the 2016 Election," *Journal of Economic Perspectives* 31, no. 2 (May 2017): 211–236.
[32] Cota et al., "Quantifying Echo Chamber Effects."
[33] Roy M. Anderson and Robert M. May, *Infectious Diseases of Humans: Dynamics and Control* (Oxford: Oxford University Press, 1992).
[34] Cinelli et al., "The Echo Chamber Effect on Social Media."
[35] Quattrociocchi, Scala, and Sunstein, *Echo Chambers on Facebook*.
[36] Michela Del Vicario et al., "The Spreading of Misinformation Online," *Proceedings of the National Academy of Sciences* 113, no. 3 (January 2016): 554–559.
[37] Vosoughi, Roy, and Aral, "Spread of True and False News Online."
[38] Andrew Guess et al., "Avoiding the Echo Chamber about Echo Chambers," *Knight Foundation* 2 (2018): 1–25.
[39] Talha Burki, "The Online Anti-Vaccine Movement in the Age of COVID-19," *Lancet Digital Health* 2, no. 10 (2020): e504–e505; "Failure to Act," Center for Countering Digital Hate, 2020, available at https://counterhate.com/research/failure-to-act/.
[40] Mønsted and Lehmann, "Characterizing Polarization in Online Vaccine Discourse"; Cinelli et al., "Echo Chamber Effect on Social Media."
[41] Cinelli et al., "Echo Chamber Effect on Social Media"; Cinelli et al., "COVID-19 Social Media Infodemic."
[42] Neil F. Johnson et al., "The Online Competition between Pro- and Anti-Vaccination Views," *Nature* 582, no. 7811 (June 2020): 230–233, https://doi.org/10.1038/s41586-020-2281-1.
[43] N55 Spohr, "Fake News and Ideological Polarization."
[44] Norbert Schwarz, Herbert Bless, Fritz Strack, Gisela Klumpp, Helga Rittenauer-Schatka, and Annette Simons, "Ease of Retrieval as Information: Another Look at the Availability Heuristic," *Journal of Personality and Social Psychology* 61, no. 2 (1991): 195; Amos Tversky and Daniel Kahneman, "Availability: A Heuristic for Judging Frequency and Probability," *Cognitive psychology* 5, no. 2 (1973): 207–232.
[45] Cass R. Sunstein, *The Law of Group Polarization*, SSRN Scholarly Paper 199668 (Rochester, NY: Social Science Research Network, December 1999), 58; Lewandowsky, Ecker, and Cook, "Beyond Misinformation."
[46] Lewandowsky, Ecker, and Cook, "Beyond Misinformation."
[47] Ángel Gómez, Lucía López-Rodríguez, Hammad Sheikh, and Jeremy Ginges et al., "The Devoted Actor's Will to Fight and the Spiritual Dimension of Human Conflict," *Nature Human Behaviour* 1, no. 9 (2017): 673–679. N59 Ángel Gómez, Lucía López-Rodríguez, Hammad Sheikh, and Jeremy Ginges, "The Devoted Actor's Will to Fight and the Spiritual Dimension of Human Conflict," *Nature Human Behaviour* 1, no. 9 (2017): 673–679.
[48] William B. Swann Jr., Ángel Gómez, D. Conor Syle, J. Francisco Morales, and Cmen Huici, "Identity Fusion: The Interplay of Personal and Social Identities in Extreme Group Behavior," *Journal of Personality and Social Psychology* 96, no. 5 (2009): 995.
[49] Dylan Weber, Ryan Theisen, and Sebastien Motsch, "Deterministic versus Stochastic Consensus Dynamics on Graphs," *Journal of Statistical Physics* 176, no. 1 (July 2019): 40–68, DOI:10.1007/s10955-019-02293-5.
[50] Rainer Hegselmann and Ulrich Krause, "Opinion Dynamics and Bounded Confidence: Models, Analysis, and Simulation," *Journal of Artificial Societies and Social Simulation* 5, no. 3 (2002).
[51] Vincent D. Blondel, Julien M. Hendrickx, and John N. Tsitsiklis, "On Krause's Multi-Agent Consensus Model with State-Dependent Connectivity," *IEEE Transactions on Automatic Control* 54, no. 11 (2009): 2586–2597; Guillaume Deffuant, David Neau, Frederic Amblard, and Gérard Weisbuch, "Mixing Beliefs among Interacting Agents," *Advances in Complex Systems* 03, no. 01n04 (January 2000): 87–98; Pierre-Emmanuel Jabin and Sabastien Motsch, "Clustering and Asymptotic Behavior in Opinion Formation," *Journal of Differential Equations* 257, no. 11 (2014): 4165–4187; Jan Lorenz, *Consensus Strikes Back in the Hegselmann Krause Model of Continuous Opinion Dynamics Under Bounded Confidence*, *Journal of Artificial Societies and Social Simulation* 9, no. 1 (January 2006); Vincent D. Blondel, Julien M. Hendrickx, and John N. Tsitsiklis, "On the 2R Conjecture for Multi-Agent Systems," in *2007 European Control Conference (ECC)* (July 2007), 874–881; Ulrich Krause, "A Discrete Nonlinear and Non-Autonomous Model of Consensus Formation," in *Communications in Difference Equations: Proceedings of the Fourth International Conference on Difference Equations* (Boca Raton, FL: CRC Press, 2000), 227.

[52] Schmidt et al., "Anatomy of News Consumption on Facebook"; Fabian Baumann, Philipp Lorenz-Spreen, Igor M. Sokolov, and Michele Starnini, "Modeling Echo Chambers and Polarization Dynamics in Social Networks," *Physical Review Letters* 124, no. 4 (January 2020): 048301; Del Vicario et al., "Spreading of Misinformation Online."

[53] Dylan Weber, GuanLin Li, and Sebastien Motsch, "Bounded Confidence Dynamics and Graph Control: Enforcing Consensus," *Networks & Heterogeneous Media* 15, no. 3 (2020).

[54] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, and Jacob Eisenstein, "You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined through Hate Speech," *Proceedings of the ACM on Human-Computer Interaction* 1, no. CSCW (2017): 1–22; Richard Rogers, "Deplatforming: Following Extreme Internet Celebrities to Telegram and Alternative Social Media," *European Journal of Communication* 35, no. 3 (2020): 213–229; Shiza Ali et al., "Understanding the Effect of Deplatforming on Social Networks," in *13th ACM Web Science Conference 2021* (2021), 187–195; Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman, "Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter," *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2 (2021): 1–30; Adrian Rauchfleisch and Jonas Kaiser, "Deplatforming the Far-Right: An Analysis of YouTube and BitChute," June 15, 2021, available at, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3867818; Helen Innes and Martin Innes, "De-Platforming Disinformation: Conspiracy Theories and Their Control," *Information, Communication & Society*, 2021, 1–19.

[55] James Madison, "The Federalist No. 10: The Same Subject (The Utility of the Union as a Safe Guard against Domestic Faction and Insurrection) Continued," November 23, 1787, in *The Federalist Papers* (New Haven: Lillian Goldman Library, Yale University, 2008), https://avalon.law.yale.edu/18th_century/fed10.asp.